Exponential Smoothing for Forecasting and Bayesian Validation of Computer Models

A Thesis
Presented to
The Academic Faculty

by

Shuchun Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Industrial and Systems Engineering
Georgia Institute of Technology
December 2006

Exponential Smoothing for Forecasting and Bayesian Validation of Computer Models

Approved by:

Professor Kwok-Leung Tsui, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Roshan J. Vengazhiyil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Wei Jiang
Department of Systems Engineering and Engineering Management
Stevens Institute of Technology

Professor Ming Yuan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor David M. Goldsman
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: August 15, 2006

To my parents

ACKNOWLEDGEMENTS

First and foremost, I would like to express my special and sincere thanks to my dissertation advisor, Dr. Kwok-Leung Tsui, for his inspiration, guidance, and encouragement during my studies in the School of Industrial and Systems Engineering at the Georgia Institute of Technology.

I would also like to express my great appreciation to Dr. David Goldsman, Dr. Wei Jiang, Dr. Roshan Joseph Vengazhiyil, and Dr. Ming Yuan for serving on my dissertation committee and for their valuable suggestions and comments. I am very grateful to Dr. Wei Jiang for his continued support, valuable discussions, and critical comments throughout my research.

I would like to extend my appreciation to all my friends at the Georgia Institute of Technology for their continued help and support.

Finally, I would like to thank my parents, my sister Shujie, my brother Shufeng, and my friend Wenzhang. Their love, support, encouragement and tolerance have helped me through difficult times and given me strength and courage to face challenges.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Despite their success and widespread usage in industry and business, ES methods have received little attention from the statistical community. We investigate three types of statistical models that have been found to underpin ES methods. They are ARIMA models, state space models with multiple sources of error (MSOE), and state space models with a single source of error (SSOE). We establish the relationship among the three classes of models and conclude that the class of SSOE state space models is broader than the other two and provides a formal statistical foundation for ES methods. To better understand ES methods, we investigate the behaviors of ES methods for time series generated from different processes. We mainly focus on time series of ARIMA type.

ES methods forecast a time series using only its own history. To include covariates into ES methods for better forecasting a time series, we propose a new forecasting method, Exponential Smoothing with Covariates (ESCov). ESCov uses an ES method to model what left unexplained in a time series by covariates. We establish the optimality of ESCov, identify SSOE state space models underlying ESCov, and derive analytically the variances of forecasts by ESCov. Empirical studies show that ESCov outperforms ES methods and regression with ARIMA errors. We suggest a model selection procedure for choosing appropriate covariates and ES methods in practice.

Computer models have been commonly used to investigate complex systems for which physical experiments are highly expensive or very time-consuming. Before using a computer model, we need to address an important question "How well does the computer model represent the real system?" The process of addressing this question is called computer model validation that generally involves the comparison of

computer outputs and physical observations. In this thesis, we propose a Bayesian approach to computer model validation. This approach integrates together computer outputs and physical observation to give a better prediction of the real system output. This prediction is then used to validate the computer model. We investigate the impacts of several factors on the performance of the proposed approach and propose a generalization to the proposed approach.

CHAPTER I

INTRODUCTION

1.1 Motivation

Industry and business have continuously used exponential smoothing (ES) methods, a collection of extrapolative forecasting methods that forecast a time series based on only its historical values. According to a mail survey conducted by Mentzer and Kahn (1995), who surveyed 207 forecasting executives, ES methods are the most common methods of forecasting. Their popularity can be attributed to several practical considerations. First, they are very simple in concept and easy to understand. Second, they require little computational effort and small data storage space. Third, they can achieve flexible adaptivity by varying smoothing parameters to account for changes in the behaviors of the time series being forecasted. Last and more importantly, ES methods, as shown by numerous empirical studies based on a wide class of time series, real or simulated (Makridakis and Hibon 1979, 2000; Makridakis et al. 1982; Makridakis et al. 1993; Chen 1997), perform as well as or sometimes better than statistically sophisticated methods such as the autoregressive integrated moving average (ARIMA) approach advocated by Box and Jenkins (Box et al. 1994). Among the various ES methods, the three best-known and most commonly used ones are simple exponential smoothing (Brown 1959, 1963), Holt's linear trend method (Holt 1957), and Holt-Winters' seasonal method (Winters 1960).

Despite their widespread usage, ES methods, initially proposed as heuristic procedures for forecasting (Brown 1959, 1963), had received little attention from the statistical community. Some statistical models such as the ARIMA models have been found to underlie certain ES methods (Muth 1960, Harrison 1967, Roberts 1982,

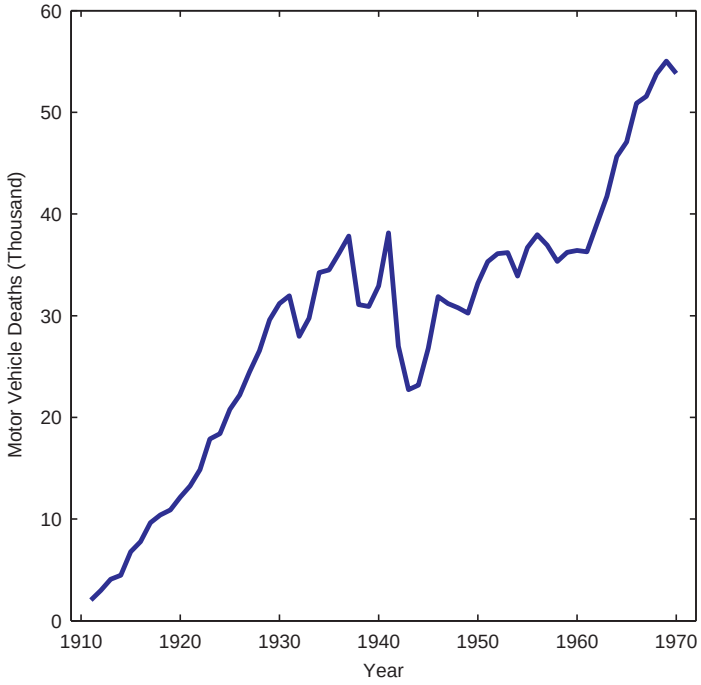
Harvey 1984, and Gardner and Mckenzie 1985). In other words, those ES methods are optimal (i.e., provide minimum mean squared error forecast) for the underlying statistical models. However, a general statistical framework was still missing. An important development in the study of ES methods occurred in 1985 when Snyder (1985) proposed a class of linear state space models that rely on only a single source of randomness and suggested the use of such models as an explanation for ES methods. However, this insight went largely unnoticed until the work of Ord et al. (1997), who expressed state space models with a single source of error in a very general form that encompasses both linear and nonlinear cases with homoscedastic or heteroscedastic variance. This general formulation provides a formal statistical foundation for ES methods. Not only is the identification of underlying statistical models for ES methods straightforward, but the underlying statistical models are also not unique. This non-uniqueness could be considered as a statistical explanation for the robustness (i.e., reasonably good performance on a wide class of time series) of ES methods.

Another way to look at the robustness of ES methods is to see how well they perform when the data are generated from a model for which they are not optimal. Cox (1961) studied the performance of simple exponential smoothing on a stationary first-order autoregressive (AR(1)) model and concluded that simple exponential smoothing performs quite well in terms of the mean squared one-step-ahead forecast error when the AR(1) model has a positive lag-one autocorrelation. Cohen (1963) examined the behaviors of simple exponential smoothing and Brown's double exponential smoothing (Brown 1963), a special case of Holt's method, when the data were generated from either a white noise process or an AR(1) model and provided a range of values of the smoothing parameters to minimize the mean squared forecast error. Cogger (1973) investigated the forecasting performance of Brown's double exponential smoothing on an ARIMA(0,1,1) model. The performance of ES methods, mainly simple exponential smoothing and Holt's method, under non-optimal situations was

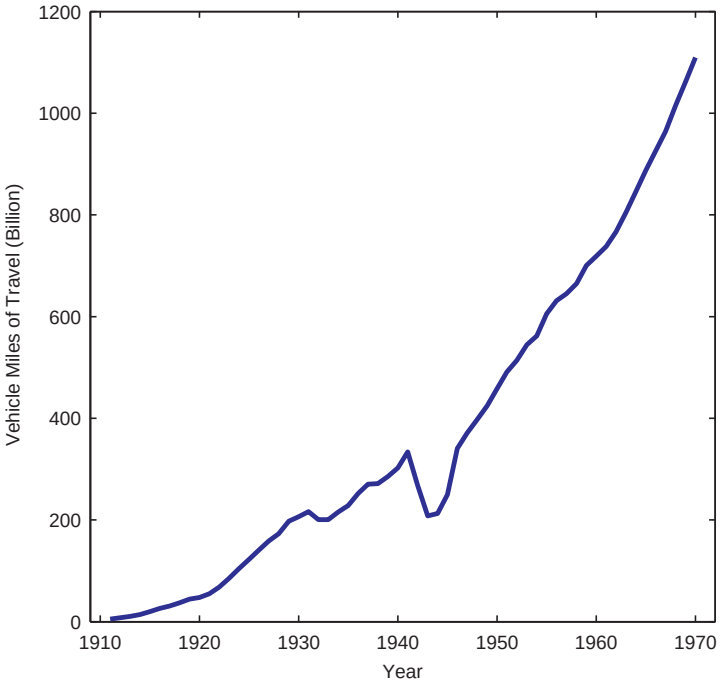
also studied in the context of process control (Ingolfsson and Sachs 1993, Castillo 2001). Nevertheless, the types of models considered were very limited.

ES methods forecast a time series using only the series' own historical values. The history of a time series certainly contains knowledge about its future. However, other information beyond what is available in a series' own history may also shed light on the series' movements along time and therefore, lead to more accurate forecasting of its future if incorporated. For example, Figure 1.1(a) displays the number of deaths (in thousands) due to motor vehicle accidents in the United States from 1911 to 1970. To forecast the motor vehicle deaths in the years after 1970, ES methods use only historical observations from 1911 to 1970. However, the number of motor vehicle deaths may be affected by various factors such as annual vehicle miles of travel, road quality, driver behaviors, and weather conditions. Figure 1.1(b) shows annual vehicle miles of travel (in billions) in the United States from 1911 to 1970. The mile series exhibits movements that appear to be correlated with those of the death series. For instance, both series dropped suddenly in 1942 and then started to climb rapidly in 1944. This suggests the use of influencing factors that explain the movements of the death series, such as fitting a regression model for the number of deaths with influencing factors as explanatory variables. Using only the series' own history for forecasting, ES methods might lose valuable information contained in influencing factors. On the other hand, influencing factors may not be able to completely explain the movements of the death series. Furthermore, although some influencing factors such as annual miles of travel are easy to measure, some influencing factors such as road quality are difficult to quantify. A possible solution would be an approach that uses ES methods to model what are left unexplained in the movements of the time series being forecasted by measurable influencing factors.

Computer models are mathematic representations of real systems, such as a group



(a)



(b)

Figure 1.1: US Motor Vehicle Deaths and Miles of Travel from 1911 to 1970

of partial differential equations with initial and boundary conditions for many engineering problems. They have been commonly used to investigate complex systems for which physical experiments are either highly expensive or too time consuming (Sacks et al 1989, Welch et al 1992, Santner et al 2003). However, before using a computer model to investigate a real system, we need to address an important question "How well does the computer model represent the real system?" Without a meaningful answer to this question, any conclusions based on the analysis of outputs from a computer model are about this computer model and can not be simply applied to the real system being investigated. The process of determining to what degree a computer model accurately represents the real system is referred to as computer model validation (AIAA G-077-1998) that generally involves the comparison of the outputs of a computer model to observations collected from physical experiments.

Recently, Oberkampf and Barone (2004) gave a comprehensive review on computer model validation. They argued that computer model validation should be done quantitatively through the use of computable measures that enable the quantitative comparison of computer outputs and physical observations. They referred to those computable measures as validation metrics. They then discussed a variety of conceptual properties that a validation metric should possess and emphasized that a validation metric should quantify uncertainties in the comparison of computer outputs and physical observations. Uncertainties could be due to random measurement errors in physical observations and/or errors resulting from post-processing computer outputs and/or physical observations, such as errors resulting from fitting a model to computer outputs and/or physical observations. In the same paper, they proposed a frequentist approach to computer model validation. Their approach for the first time validates a computer model by quantitatively comparing computer outputs and physical observations over a range of input variables. In this thesis, we propose a Bayesian approach to computer model validation. The Bayesian approach has the

ability to take into consideration prior knowledge on the real system in the form of prior distributions for certain parameters. It outputs the posterior distributions of both the model bias (defined as the difference of the computer model output and the real system output) and the real system output given computer outputs and physical observations. The posterior distribution of the model bias serves as a validation metric for the quantitative comparison of computer outputs and physical observations. Both the mean and variance of the model bias are functions of input variables, providing quantitative measures of the representativeness of the computer model over a range of input variables. The posterior distribution of the real system output provides a more accurate prediction of the real system output.

1.2 Organization

This thesis consists of two parts. The first part, spanning Chapters two, three, and four, concerns the statistical properties of ES methods and the incorporation of explanatory variables into ES forecasting. The second part, consisting of Chapters five and six, covers the topic on computer model validation. The organization of this thesis is as follows. Chapter one explains the motivations. Chapter two investigates statistical models underlying ES methods and provides a general statistical framework for ES methods. Chapter three discusses the robustness of ES methods by examining their performance for different ARIMA-type data generating processes. Chapter four proposes a new forecasting method that bases forecasts on not only the time series being forecasted but also explanatory variables affecting the movements of the series and studies its performance and statistical properties. Chapter five reviews current practices in computer model validation and proposed a Bayesian approach to the validation of computer models. Chapter six investigates the performance of the proposed approach in different situations and proposes a possible generalization to the proposed approach.

CHAPTER II

STATISTICAL FOUNDATIONS OF EXPONENTIAL SMOOTHING (ES) METHODS

2.1 Introduction

Widely used in industry and commerce for forecasting, ES methods are often considered as *ad hoc* procedures with no formal statistical foundation. Contrary to this general-accepted but outdated perception, different types of statistical models have been found to underpin ES methods. Moreover, such underpinning statistical models, as shown later, are not unique. This non-uniqueness helps to explain the robustness of ES methods. This chapter starts with a brief review on ES methods and an introduction to three types of statistical models that have been found to underlie ES methods. Then, possible underlying models are identified for various ES methods, followed by a discussion on the relationships among those three types of underlying statistical models.

As regard notations, let Y_t denote the observed value of a time series at time t, $\hat{Y}_{t|t-h}$ the prediction (or forecast) of Y_t made at time $t-h$, and $e_{t|t-h}$ the corresponding forecast error

$$e_{t|t-h} = Y_t - \hat{Y}_{t|t-h},\ t > 0,\ h > 0 \tag{2.1}$$

When $h = 1$, e_t is used in place of $e_{t|t-1}$ for simplicity as well as for consistency with the literature.

2.2 ES methods

This section reviews ES methods. Following a detailed description of the three best-known ES methods, simple exponential smoothing, Holt's linear trend method, and Holt-Winters' seasonal method, is a discussion on the developments of other less common methods. This section ends with a taxonomy of ES methods.

2.2.1 Simple Exponential Smoothing (SES)

A time series with a local constant level can be adequately represented by a model of the form

$$Y_t = \mu_t + \epsilon_t, \tag{2.2}$$

where ϵ_t is a white noise process with $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = \sigma^2$. The level μ_t may change slowly with time. However, in any local time segment, a constant μ gives a reasonably good model of the time series.

Let l_t denote the estimator of the level at time t. Given l_{t-1}, once the observation at time t, Y_t, becomes available, SES due to Brown (1959, 1963) updates the level estimator via the recurrence equation

$$l_t = \alpha Y_t + (1 - \alpha) l_{t-1}, \tag{2.3}$$

where α is a smoothing parameter taking values in the interval $(0, 1]$. According to equation (2.3), the estimator at time t, l_t, is a weighted average of the latest observation, Y_t, and the estimator at time $t-1$, l_{t-1}. The value of α can be used to adjust the sensitivity of the estimator to changes in the series level. The larger the α value, the higher weight Y_t receives, the more sensitive the estimator to changes in the level. The h-step-ahead forecast made at time t by SES is

$$\hat{Y}_{t+h|t} = l_t, \ h > 0, \tag{2.4}$$

and the corresponding h-step-ahead forecast error is

$$e_{t+h|t} = Y_{t+h} - \hat{Y}_{t+h|t} = Y_{t+h} - l_t. \tag{2.5}$$

Recurrence equation (2.3) has an equivalent error-correction form

$$l_t = l_{t-1} + \alpha e_t, \tag{2.6}$$

where e_t (rigorously, e_t should be written as $e_{t|t-1}$) is the one-step-ahead forecast error

$$e_t = Y_t - \hat{Y}_{t|t-1} = Y_t - l_{t-1}. \tag{2.7}$$

By successive substitution, equation (2.6) can be rewritten as a weighted average of all past observations with weights declining exponentially

$$l_t = \alpha \sum_{i=0}^{\infty} (1-\alpha)^i Y_{t-i}, \tag{2.8}$$

thus the name *exponential smoothing*.

2.2.2 Holt's Method

SES does not perform well for forecasting time series with a local linear trend, namely,

$$Y_t = \mu_t + \epsilon_t = \beta_0 + \beta_1 t + \epsilon_t, \tag{2.9}$$

where β_0 and β_1 are assumed to be constant in any local time segment, but may change slowly with time. It can be shown that the forecast by SES for a time series with a local linear trend tends to fall behind the series itself (Brown 1963). To better forecast time series with a local linear trend, Holt (1957) extended SES by adding one more updating equation for the slope β_1.

Let b_t denote the estimator of the slope at time t. Given l_{t-1} and b_{t-1}, once the observation at time t, Y_t, becomes available, Holt's method updates the estimators using the recurrence equations

$$l_t = \alpha_1 Y_t + (1-\alpha_1)(l_{t-1} + b_{t-1}), \tag{2.10a}$$

$$b_t = \alpha_2 (l_t - l_{t-1}) + (1-\alpha_2) b_{t-1}, \tag{2.10b}$$

where α_1 and α_2 are smoothing parameters taking values in the interval $(0, 1]$. The h-step-ahead forecast made at time t by Holt's method is

$$\hat{Y}_{t+h|t} = l_t + h b_t, \ h > 0, \tag{2.11}$$

and the corresponding h-step-ahead forecast error is

$$e_{t+h|t} = Y_{t+h} - \hat{Y}_{t+h|t} = Y_{t+h} - l_t - hb_t. \tag{2.12}$$

Similar to SES, Holt's method can be written in an equivalent error-correction form (Gardner 1985)

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t, \tag{2.13a}$$

$$b_t = b_{t-1} + \alpha_1\alpha_2 e_t, \tag{2.13b}$$

where $e_t = Y_t - l_{t-1} - b_{t-1}$ is the one-step-ahead forecast error.

Brown (1963) also proposed a local linear trend forecasting procedure called double exponential smoothing, which uses a single smoothing parameter to smooth both the level and the slope. Brown's double exponential smoothing, in error-correction form, is given by

$$l_t = l_{t-1} + b_{t-1} + \alpha(2-\alpha)e_t, \tag{2.14a}$$

$$b_t = b_{t-1} + \alpha^2 e_t, \tag{2.14b}$$

Comparing equations (2.14a) and (2.14b) to (2.13a) and (2.13b) reveals that Brown's double exponential smoothing with a smoothing parameter α is equivalent to Holt's method with smoothing parameters $\alpha_1 = \alpha(2-\alpha)$ and $\alpha_2 = \alpha/(2-\alpha)$. That is, Brown's double exponential method is a special case of Holt's method.

2.2.3 Holt-Winters' Method

Neither SES nor Holt's local linear method are appropriate for forecasting time series with seasonal changes. To capture seasonal changes, Winters (1960) generalized Holt's local linear method and proposed the so-called Holt-Winters' method, which has three updating equations, one for the level, one for the slope, and one for the seasonality. Depending on whether the seasonality is combined with the linear trend additively or multiplicatively, there are two versions of Holt-Winters' method.

• **Additive Holt-Winters' Method**

A time series with a local linear trend and an additive seasonality can be represented by a model of the form

$$Y_t = \beta_0 + \beta_1 t + s_t + \epsilon_t, \tag{2.15}$$

where s_t is the seasonal factor at time t.

Let c_t denote the estimator of s_t. The updating equations of the additive Holt-Winters' method are given by, in error-correction form,

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t, \tag{2.16a}$$

$$b_t = b_{t-1} + \alpha_1 \alpha_2 e_t, \tag{2.16b}$$

$$c_t = c_{t-M} + (1 - \alpha_1)\alpha_3 e_t, \tag{2.16c}$$

where M is the length of a complete seasonal cycle, and α_1, α_2, and α_3 are smoothing parameters taking values in the interval (0,1]. The h-step-ahead forecast made at time t by the additive Holt-Winters' method is

$$\hat{Y}_{t+h|t} = l_t + hb_t + c_{t-M+h},\ h > 0. \tag{2.17}$$

• **Multiplicative Holt-Winters' Method**

A time series with a local linear trend and a multiplicative seasonality can be represented by a model of the form

$$Y_t = (\beta_0 + \beta_1 t) \cdot s_t + \epsilon_t. \tag{2.18}$$

The updating equations for the multiplicative Holt-Winters' method are given by, in error-correction form,

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t / c_{t-M}, \tag{2.19a}$$

$$b_t = b_{t-1} + \alpha_1 \alpha_2 e_t / c_{t-M}, \tag{2.19b}$$

$$c_t = c_{t-M} + (1 - \alpha_1)\alpha_3 e_t / l_t. \tag{2.19c}$$

The h-step-ahead forecast made at time t by the multiplicative Holt-Winters' method is

$$\hat{Y}_{t+h|t} = (l_t + hb_t) \cdot c_{t-M+h},\ h > 0. \tag{2.20}$$

2.2.4 Other ES methods

Many other ES methods have been proposed for time series forecasting. Empirical studies have shown that Holt's linear trend method tends to overshot the data when used for long-term forecasting. To overcome this problem, Gardner and Mckenzie (1985) introduced an extra parameter, a dampening parameter ϕ $(0 \leq \phi \leq 1)$, into Holt's method to gain more controls over trend extrapolations. The resulting method is referred to as damped Holt's method. Believing that real-life series more likely have a multiplicative trend than an additive one, Pegels (1969) proposed a smoothing method for forecasting time series with a multiplicative trend. Pegels' method smoothes the ratio l_t/l_{t-1}, instead of difference $l_t - l_{t-1}$ as in Holt's method, of successive series levels. Taylor (2003), in an analogous way that Holt's linear trend method is damped, added a dampening parameter ϕ to Pegels' method for forecasting time series with a damped multiplicative trend. ES methods for polynomial trends of order $k, k > 1$, were also developed (Gardner 1985, Montgomery, Johnson and Gardiner 1990). However, their usages have been discouraged both from practical considerations and by empirical studies (Makridakis et al. 1982).

Pegels (1969) proposed a taxonomy of ES methods, which was extended and modified later by Gardner and Mckenzie (1985), Hyndman et al. (2002), and Taylor (2003). This taxonomy is presented in Table 2.1. In this table, each method has two components, a trend and a seasonality. For example, N-N represents the ES method with neither trend nor seasonality, namely SES; A-N represents the method with an additive trend but no seasonality, namely Holt's method; and A-A and A-M represent the additive and multiplicative Holt-Winters' methods respectively.

Table 2.2 gives the updating equations in error-correction form as well as the h-step-ahead forecast for all of the methods listed in Table 2.1. The updating equations for the seasonal factor c_t in multiplicative seasonality cases are slightly different from the commonly used ones. l_{t-1} as in N-M (or $l_{t-1} + b_{t-1}$ as in A-M or $l_{t-1} + \phi b_{t-1}$ as in DA-M) in place of l_t is used in the denominator. This small modification is convenient to the identification of underlying single source of error state space models later. Although their derivations are straightforward, the formulas for DM-A and DM-M have not, to our knowledge, explicitly appeared in print before.

Table 2.1: ES methods

	Seasonality		
Trend	None	Additive	Multiplicative
None	N-N	N-A	N-M
Additive	A-N	A-A	A-M
Multiplicative	M-N	M-A	M-M
Damped Additive	DA-N	DA-A	DA-M
Damped Multiplicative	DM-N	DM-A	DM-M

2.3 Three Statistical Models

Three types of statistical models have been found to underpin ES methods. They are autoregressive integrated moving average (ARIMA) models, state space models with multiple sources of error, and state space models with single source of error.

2.3.1 ARIMA Model

ARIMA models and their applications in time series forecasting are discussed in many textbooks (Box et al. 1994, Chatfield 1996, Blackwell and Davis 1991). Here we only gives a very general formula of ARIMA models. For more detailed information, such as model identification, estimation, and forecasting, please refer to the books mentioned above.

Table 2.2: Error-Correction Form Updating Equations and h-step-ahead Forecasts of ES methods (**N** - None, **A** - Additive, **M** - Multiplicative, **DA** - Damped Additive, **DM** - Damped Multiplicative)

	Seasonality		
Trend	**N**	**A**	**M**
N	$l_t = l_{t-1} + \alpha_1 e_t$	$l_t = l_{t-1} + \alpha_1 e_t$	$l_t = l_{t-1} + \alpha_1 e_t / c_{t-M}$
		$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t$	$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t / l_{t-1}$
	$\hat{Y}_{t+h\|t} = l_t$	$\hat{Y}_{t+h\|t} = l_t + c_{t-M+h}$	$\hat{Y}_{t+h\|t} = l_t c_{t-M+h}$
A	$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t$	$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t$	$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t / c_{t-M}$
	$b_t = b_{t-1} + \alpha_1\alpha_2 e_t$	$b_t = b_{t-1} + \alpha_1\alpha_2 e_t$	$b_t = b_{t-1} + \alpha_1\alpha_2 e_t / c_{t-M}$
		$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t$	$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t / (l_{t-1} + b_{t-1})$
	$\hat{Y}_{t+h\|t} = l_t + hb_t$	$\hat{Y}_{t+h\|t} = l_t + hb_t + c_{t-M+h}$	$\hat{Y}_{t+h\|t} = (l_t + hb_t) c_{t-M+h}$
M	$l_t = l_{t-1} b_{t-1} + \alpha_1 e_t$	$l_t = l_{t-1} b_{t-1} + \alpha_1 e_t$	$l_t = l_{t-1} b_{t-1} + \alpha_1 e_t / c_{t-M}$
	$b_t = b_{t-1} + \alpha_1\alpha_2 e_t / l_{t-1}$	$b_t = b_{t-1} + \alpha_1\alpha_2 e_t / l_{t-1}$	$b_t = b_{t-1} + \alpha_1\alpha_2 e_t / (l_{t-1} c_{t-M})$
		$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t$	$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t / (l_{t-1} b_{t-1})$
	$\hat{Y}_{t+h\|t} = l_t b_t^h$	$\hat{Y}_{t+h\|t} = l_t b_t^h + c_{t-M+h}$	$\hat{Y}_{t+h\|t} = (l_t b_t^h) c_{t-M+h}$
DA	$l_t = l_{t-1} + \phi b_{t-1} + \alpha_1 e_t$	$l_t = l_{t-1} + \phi b_{t-1} + \alpha_1 e_t$	$l_t = l_{t-1} + \phi b_{t-1} + \alpha_1 e_t / c_{t-M}$
	$b_t = \phi b_{t-1} + \alpha_1\alpha_2 e_t$	$b_t = \phi b_{t-1} + \alpha_1\alpha_2 e_t$	$b_t = \phi b_{t-1} + \alpha_1\alpha_2 e_t / c_{t-M}$
		$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t$	$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t / (l_{t-1} + \phi b_{t-1})$
	$\hat{Y}_{t+h\|t} = l_t + b_t \sum_{i=1}^{h} \phi^i$	$\hat{Y}_{t+h\|t} = l_t + b_t \sum_{i=1}^{h} \phi^i + c_{t-M+h}$	$\hat{Y}_{t+h\|t} = (l_t + b_t \sum_{i=1}^{h} \phi^i) c_{t-M+h}$
DM	$l_t = l_{t-1} b_{t-1}^{\phi} + \alpha_1 e_t$	$l_t = l_{t-1} b_{t-1}^{\phi} + \alpha_1 e_t$	$l_t = l_{t-1} b_{t-1}^{\phi} + \alpha_1 e_t / c_{t-M}$
	$b_t = b_{t-1}^{\phi} + \alpha_1\alpha_2 e_t / l_{t-1}$	$b_t = b_{t-1}^{\phi} + \alpha_1\alpha_2 e_t / l_{t-1}$	$b_t = b_{t-1}^{\phi} + \alpha_1\alpha_2 e_t / (l_{t-1} c_{t-M})$
		$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t$	$c_t = c_{t-M} + (1-\alpha_1)\alpha_3 e_t / (l_{t-1} b_{t-1}^{\phi})$
	$\hat{Y}_{t+h\|t} = l_t b_t^{\sum_{i=1}^{h} \phi^i}$	$\hat{Y}_{t+h\|t} = l_t b_t^{\sum_{i=1}^{h} \phi^i} + c_{t-M+h}$	$\hat{Y}_{t+h\|t} = (l_t b_t^{\sum_{i=1}^{h} \phi^i}) c_{t-M+h}$

ARIMA models can be written in a very general form

$$(1-\sum_{i=1}^{p}\phi_i B^i-\sum_{i=1}^{P}\Phi_i B^{iM})(1-B)^d(1-B^M)^D Y_t=(1+\sum_{i=1}^{q}\theta_i B^i+\sum_{i=1}^{P}\Theta_i B^{iM})\epsilon_t \tag{2.21}$$

for an additive seasonality, or

$$(1-\sum_{i=1}^{p}\phi_i B^i)(1-\sum_{i=1}^{P}\Phi_i B^{iM})(1-B)^d(1-B^M)^D Y_t=(1+\sum_{i=1}^{q}\theta_i B^i)(1+\sum_{i=1}^{P}\Theta_i B^{iM})\epsilon_t \tag{2.22}$$

for a multiplicative seasonality, where ϵ_t is a white noise process with $E[\epsilon_t]=0$ and $E[\epsilon_t^2]=\sigma^2$; B is the back shift operator; $\phi_i, 1\le i\le p$, and $\theta_i, 1\le i\le q$, are non-seasonal ARMA parameters; $\Phi_i, 1\le i\le P$, and $\Theta_i, 1\le i\le Q$, are seasonal ARMA parameters; p and q are non-seasonal AR and MA orders respectively; P and Q are seasonal AR and MA orders respectively; d is the order of non-seasonal differencing; D is the order of seasonal differencing; and M is the length of a complete seasonal cycle. Models (2.21) and (2.22) have a short representation

$$\text{ARIMA}(p,d,q)\oplus\text{ARIMA}(P,D,Q)_M \tag{2.23}$$

with $\oplus$ being "+" for additive seasonal ARIMA models and "×" for multiplicative seasonal ARIMA models. Non-seasonal ARIMA models have an even shorter representation

$$\text{ARIMA}(p,d,q). \tag{2.24}$$

We only concern these ARIMA models that are stationary and invertible. Stationarity requires that the roots of

$$1-\sum_{i=1}^{p}\phi_i B^i-\sum_{i=1}^{P}\Phi_i B^{iM}=0 \tag{2.25}$$

or

$$(1-\sum_{i=1}^{p}\phi_i B^i)(1-\sum_{i=1}^{P}\Phi_i B^{iM})=0 \tag{2.26}$$

have an absolute value great than 1. Invertibility requires that the roots of

$$1 + \sum_{i=1}^{q} \theta_i B^i + \sum_{i=1}^{P} \Theta_i B^{iM} = 0 \tag{2.27}$$

or

$$(1 + \sum_{i=1}^{q} \theta_i B^i)(1 + \sum_{i=1}^{P} \Theta_i B^{iM}) = 0 \tag{2.28}$$

are greater than 1 in absolute value. For instance, the invertibility condition for an ARIMA(0,1,1) model is

$$-1 < \theta_1 < -1 \tag{2.29}$$

while the invertibility conditions for an ARIMA(0,2,2) model are

$$-(1 + \theta_2) < \theta_1 < 1 + \theta_2 \text{ and } -1 < \theta_2 < 1. \tag{2.30}$$

2.3.2 Multiple Source of Error (MSOE) State Space Model

MSOE State space models consist of two equations, an observation equation

$$Y_t = \mathbf{x}_t^T \boldsymbol{\beta}_t + \epsilon_t \tag{2.31}$$

and a transition equation

$$\boldsymbol{\beta}_t = \mathbf{G}_t \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \tag{2.32}$$

where $\mathbf{x}_t$ is a $p \times 1$ covariate vector, and $\mathbf{G}_t$ is a $p \times p$ transition matrix. Both $\mathbf{x}_t$ and $\mathbf{G}_t$ are assumed to be known. $\boldsymbol{\beta}_t$ is a $p \times 1$ state vector. ϵ_t and $\boldsymbol{\eta}_t$ are mutually independent white noise processes with zero mean and $E[\epsilon_t^2] = \sigma_t^2$ and $E[\boldsymbol{\eta}_t \boldsymbol{\eta}_t^T] = \mathbf{Q}_t$. ϵ_t and $\boldsymbol{\eta}_t$ might be referred to as the transient error and the permanent error respectively as ϵ_t affects only the current observation Y_t while the effect of $\boldsymbol{\eta}_t$ will persist through time. For a model in a state space form as given above, a recursive procedure, Kalman Filter (KF), can be used to calculate optimal predictions of future observations and optimal estimators of the state vector (Kalman 1960, Duncan and Horn 1972).

Let $\hat{\boldsymbol{\beta}}_{t|k}$ denote the estimator of the state vector $\boldsymbol{\beta}_t$ based on all the information up to and including time k. Let $\mathbf{P}_{t|k}$ denote the $p \times p$ covariance matrix of the associated

estimation error, $\hat{\boldsymbol{\beta}}_{t|k} - \boldsymbol{\beta}_t$. That is,

$$\mathbf{P}_{t|k} = E\left[(\hat{\boldsymbol{\beta}}_{t|k} - \boldsymbol{\beta}_t)(\hat{\boldsymbol{\beta}}_{t|k} - \boldsymbol{\beta}_t)^T\right], \tag{2.33}$$

which is also referred to as the mean squared error (MSE) matrix of $\hat{\boldsymbol{\beta}}_{t|k}$. To simplify, we use $\hat{\boldsymbol{\beta}}_t$ and $\mathbf{P}_t$ instead of $\hat{\boldsymbol{\beta}}_{t|t}$ and $\mathbf{P}_{t|t}$ when $k = t$.

At time $t-1$, given $\hat{\boldsymbol{\beta}}_{t-1}$ and $\mathbf{P}_{t-1}$, the estimator of $\boldsymbol{\beta}_t$ and the corresponding MSE matrix are given by the prediction equations

$$\hat{\boldsymbol{\beta}}_{t|t-1} = \mathbf{G}_t \hat{\boldsymbol{\beta}}_{t-1}, \tag{2.34a}$$

$$\mathbf{P}_{t|t-1} = \mathbf{G}_t \mathbf{P}_{t-1} \mathbf{G}_t^T + \mathbf{Q}_t, \tag{2.34b}$$

from which we can calulate the one-step-ahead prediction of Y_t

$$\hat{Y}_{t|t-1} = \mathbf{x}_t^T \hat{\boldsymbol{\beta}}_{t|t-1} \tag{2.35}$$

and the corresponding one-step-ahead prediction error

$$e_t = Y_t - \hat{Y}_{t|t-1} = Y_t - \mathbf{x}_t^T \hat{\boldsymbol{\beta}}_{t|t-1}. \tag{2.36}$$

Once the observation at time t, Y_t, becomes available, the estimator of $\boldsymbol{\beta}_t$ and its MSE matrix are updated via

$$\hat{\boldsymbol{\beta}}_t = \hat{\boldsymbol{\beta}}_{t|t-1} + \frac{\mathbf{P}_{t|t-1}\mathbf{x}_t}{\mathbf{x}_t^T \mathbf{P}_{t|t-1}\mathbf{x}_t + \sigma_t^2} \cdot (Y_t - \mathbf{x}_t^T \hat{\boldsymbol{\beta}}_{t|t-1}), \tag{2.37a}$$

$$\mathbf{P}_t = \mathbf{P}_{t|t-1} - \frac{\mathbf{P}_{t|t-1}\mathbf{x}_t\mathbf{x}_t^T\mathbf{P}_{t|t-1}}{\mathbf{x}_t^T \mathbf{P}_{t|t-1}\mathbf{x}_t + \sigma_t^2}. \tag{2.37b}$$

Under the assumptions that the two white noise processes $\{\epsilon_t\}$ and $\{\boldsymbol{\eta}_t\}$ are Gaussian processes, and the initial state $\boldsymbol{\beta}_0$ has a multivariate normal distribution with mean $\hat{\boldsymbol{\beta}}_0$ and covariance matrix $\mathbf{P}_0$, the updating equations (2.37a) and (2.37b) can be easily derived using the Bayesian theory (Meinhold and Singpurwalla, 1983) and $\hat{\boldsymbol{\beta}}_t$ simply is the posterior mean of $\boldsymbol{\beta}_t | Y_t, Y_{t-1}, \cdots, Y_1$. According to equation (2.37a), $\hat{\boldsymbol{\beta}}_t$ is a linear combination of the observations $Y_t, \cdots, Y_1$, therefore a linear estimator

of $\boldsymbol{\beta}_t$. $\hat{\boldsymbol{\beta}}_t$ is also known to be the minimum mean squared error (MMSE) estimator of $\boldsymbol{\beta}_t$ under the normality assumptions (Harvey 1990). Without the normality assumptions, $\hat{\boldsymbol{\beta}}_t$ is the MMSE estimator of $\boldsymbol{\beta}_t$ based on all observations up to and including time t within the class of all linear estimators (Duncan and Horn 1972).

2.3.3 Single Source of Error (SSOE) State Space Model

The MSOE state space models in equations (2.31) and (2.32) involve two mutually independent random errors, the transient error ϵ_t in the observation equation and the permanent error $\boldsymbol{\eta}_t$ in the transition equation. An alternative to the independence assumption is to assume that ϵ_t and $\boldsymbol{\eta}_t$ are perfectly correlated

$$\boldsymbol{\eta}_t = \boldsymbol{\rho}_t \epsilon_t, \tag{2.38}$$

where $\boldsymbol{\rho}_t$ is a $p \times 1$ vector. Under this assumption, the state space model in equations (2.31) and (2.32) becomes

$$Y_t = \mathbf{x}_t^T \boldsymbol{\beta}_t + \epsilon_t, \tag{2.39a}$$

$$\boldsymbol{\beta}_t = \mathbf{G}_t \boldsymbol{\beta}_{t-1} + \boldsymbol{\rho}_t \epsilon_t, \tag{2.39b}$$

which has only one source of error, namely ϵ_t, and is referred to as a single source of error (SSOE) state space model. Comparing with the MSOE model in equations (2.31) and (2.32), the SSOE model in (2.39a) and (2.39b) requires the specification of a $p \times 1$ vector $\boldsymbol{\rho}_t$ rather than a $p \times p$ covariance matrix $\mathbf{Q}_t$.

Substituting equation (2.39b) into equation (2.39a) gives

$$Y_t = \mathbf{x}_t^T \mathbf{G}_t \boldsymbol{\beta}_{t-1} + (\mathbf{x}_t^T \boldsymbol{\rho}_t + 1)\epsilon_t. \tag{2.40}$$

Let $\mathbf{z}_t = \mathbf{x}_t^T \mathbf{G}_t$, $\xi_t = (\mathbf{x}_t^T \boldsymbol{\rho}_t + 1)\epsilon_t$, and $\boldsymbol{\alpha}_t = \boldsymbol{\rho}_t / (\mathbf{x}_t^T \boldsymbol{\rho}_t + 1)$, we have the more traditional form of SSOE state space models (Ord et al. 1997)

$$Y_t = \mathbf{z}_t^T \boldsymbol{\beta}_{t-1} + \xi_t, \tag{2.41a}$$

$$\boldsymbol{\beta}_t = \mathbf{G}_t \boldsymbol{\beta}_{t-1} + \boldsymbol{\alpha}_t \xi_t, \tag{2.41b}$$

in which the observation equation involves $\boldsymbol{\beta}_{t-1}$ rather than $\boldsymbol{\beta}_t$.

Estimation and prediction for SSOE state space models can be carried out easily. In equation (2.41a), $\mathbf{z}_t^T\boldsymbol{\beta}_{t-1}$ is the one-step-ahead prediction of Y_t made at time $t-1$, and ξ_t is the corresponding one-step-ahead prediction error. Solving equation (2.41a) for ξ_t results in $\xi_t = Y_t - \mathbf{z}_t^T\boldsymbol{\beta}_{t-1}$. Substituting it into equation (2.41b) gives the updating equation for the state vector $\boldsymbol{\beta}_t$

$$\boldsymbol{\beta}_t = \mathbf{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\alpha}_t(Y_t - \mathbf{z}_t^T\boldsymbol{\beta}_{t-1}). \tag{2.42}$$

Snyder (1985) is the first one discussing SSOE state space models. His model is slightly different from the model in equations (2.39a) and (2.39b) in that, in his model, the transition equation involves ϵ_{t-1} instead of ϵ_t

$$\boldsymbol{\beta}_t = \mathbf{G}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\rho}_t\epsilon_{t-1}. \tag{2.43}$$

Snyder (1985) only investigated SSOE state space models that are linear and homoscedastic (i.e., $E[\epsilon_t^2]$ is constant). Ord et al. (1997) generalized SSOE models for nonlinear and heteroscedastic cases.

2.4 Statistical Models Underlying ES methods

This section investigates underlying statistical models for various ES methods.

2.4.1 ARIMA Model

It is well known that (Muth 1960, Roberts 1982)

a). SES with the smoothing parameter

$$\alpha = 1 + \theta_1 \tag{2.44}$$

provides optimal forecast for the ARIMA(0,1,1) model

$$(1-B)Y_t = (1+\theta_1 B)\epsilon_t, \tag{2.45a}$$

$$-1 < \theta_1 \leq 0; \tag{2.45b}$$

b). Holt's method with α_1 and α_2 given by

$$\alpha_1 = 1 - \theta_2 \quad \text{and} \quad \alpha_2 = \frac{1 + \theta_1 + \theta_2}{1 - \theta_2} \tag{2.46}$$

provides optimal forecast for the ARIMA(0,2,2) model

$$(1 - B)^2 Y_t = (1 + \theta_1 B + \theta_2 B^2)\epsilon_t, \tag{2.47a}$$

$$-(1 + \theta_2) < \theta_1 \leq -2\theta_2, \ 0 \leq \theta_2 < 1. \tag{2.47b}$$

The more restricted conditions on θ_1 and θ_2 in equations (2.45b) and (2.47b) than these imposed by invertibility (see equations (2.29) and (2.30)) are due to the fact that the smoothing parameters in ES methods are confined to the interval (0,1]. Smoothing parameters with values greater than 1 might be used. However, they are hard to explain and then not considered here. For example, for SES, when $1 < \alpha < 2$, equation (2.8) shows that weights assigned to past observations will oscillate in sign, which is counterintuitive and very unusual in real-world applications.

It was shown that Brown's double exponential smoothing with a smoothing parameter α is equivalent to Holt's method with smoothing parameters $\alpha_1 = \alpha(2 - \alpha)$ and $\alpha_2 = \alpha/(2 - \alpha)$, at which equations (2.46) and (2.47) become

$$\alpha = 1 - \sqrt{\theta_2} \tag{2.48}$$

and

$$(1 - B)^2 Y_t = (1 - \sqrt{\theta_2} B)^2 \epsilon_t, \tag{2.49a}$$

$$0 \leq \theta_2 < 1. \tag{2.49b}$$

That is, Brown's double exponential smoothing with $\alpha = 1 - \sqrt{\theta_2}$ provides optimal forecast for the equal-root ARIMA(0,2,2) model in equations (2.49a) and (2.49b). Furthermore, as Brown's double exponential smoothing is a special case of Holt's method, this equal-root ARIMA(0,2,2) model is a subclass of the ARIMA(0,2,2) model in (2.47a) and (2.47b).

Roberts (1982) and Gardner and Mckenzie (1985) also identified the ARIMA models that underpin the damped Holt's method and the additive Holt-Winters' method. In fact, all of the additive ES methods in Table 2.1 (i.e., ES methods without multiplicative components, namely N-N, N-A, A-N, A-A, DA-N, and DA-A) have underlying ARIMA models. For example, for DA-A, replacing the one-step-ahead forecast $\hat{Y}_{t|t-1}$ by $Y_t - e_t$ yields

$$Y_t = l_{t-1} + \phi b_{t-1} + c_{t-M} + e_t, \tag{2.50}$$

where $\{e_t, t > 0\}$ is a sequence of zero mean, uncorrelated one-step-ahead forecast errors. Introducing the back shift operator B into the updating equations gives

$$(1 - B)l_t = \phi b_{t-1} + \alpha_1 e_t, \tag{2.51a}$$

$$(1 - \phi B)b_t = \alpha_1\alpha_2 e_t, \tag{2.51b}$$

$$(1 - B^M)c_t = (1 - \alpha_1)\alpha_3 e_t. \tag{2.51c}$$

Left multiplying equation (2.50) by $(1-\phi B)(1-B^M)$ on both sides and substituting equation (2.51) into the result, we have the ARIMA model underlying DA-A

$$\begin{aligned}
&(1 - \phi B)(1 - B^M)Y_t = (1 + \sum_{i=1}^{M+1} \theta_i B^i)e_t,\\
&\theta_1 = \alpha_1 + \phi\alpha_1\alpha_2 - \phi,\\
&\theta_i = (1 - \phi)\alpha_1 + \phi\alpha_1\alpha_2, \quad i = 2, \cdots, M-1,\\
&\theta_M = (1 - \phi)\alpha_1 + \phi\alpha_1\alpha_2 + (1 - \alpha_1)\alpha_3 - 1,\\
&\theta_{M+1} = \phi(1 - \alpha_1) - \phi(1 - \alpha_1)\alpha_3,
\end{aligned}$$

This ARIMA model, when $\phi = 1$, becomes the ARIMA model underlying A-A; and, when $\phi = 0$, becomes the ARIMA model underlying N-A.

Table 2.3 gives the underlying ARIMA models for each additive ES method in Table 2.1 and the corresponding relationships between the ARIMA parameters and the smoothing parameters.

Table 2.3: Underlying ARIMA Models for ES methods (**N** - None, **A** - Additive, **M** - Multiplicative, **DA** - Damped Additive, **DM** - Damped Multiplicative)

	Seasonality		
Trend	**N**	**A**	**M**
N	$(1-B)Y_t = (1+\theta_1 B)\epsilon_t$ $\theta_1 = \alpha_1 - 1$	$(1-B^M)Y_t = (1+\sum_{i=1}^{M}\theta_i B^i)\epsilon_t$ $\theta_i = \alpha_1, \quad i = 1, \cdots, M-1$ $\theta_M = \alpha_1 + (1-\alpha_1)\alpha_3 - 1$	-
A	$(1-B)^2 Y_t = (1+\theta_1 B + \theta_2 B^2)\epsilon_t$ $\theta_1 = \alpha_1 + \alpha_1\alpha_2 - 2$ $\theta_2 = 1 - \alpha_1$	$(1-B)(1-B^M)Y_t = (1+\sum_{i=1}^{M+1}\theta_i B^i)\epsilon_t$ $\theta_1 = \alpha_1 + \alpha_1\alpha_2 - 1$ $\theta_i = \alpha_1\alpha_2, \quad i = 2, \cdots, M-1$ $\theta_M = \alpha_1\alpha_2 + (1-\alpha_1)\alpha_3 - 1$ $\theta_{M+1} = 1 - \alpha_1 - (1-\alpha_1)\alpha_3$	-
M	-	-	-
DA	$(1-\phi B)(1-B)Y_t = (1+\theta_1 B + \theta_2 B^2)\epsilon_t$ $\theta_1 = \alpha_1 + \phi\alpha_1\alpha_2 - \phi - 1$ $\theta_2 = \phi(1-\alpha_1)$	$(1-\phi B)(1-B^M)Y_t = (1+\sum_{i=1}^{M+1}\theta_i B^i)\epsilon_t$ $\theta_1 = \alpha_1 + \phi\alpha_1\alpha_2 - \phi$ $\theta_i = (1-\phi)\alpha_1 + \phi\alpha_1\alpha_2, \quad i = 2, \cdots, M-1$ $\theta_M = (1-\phi)\alpha_1 + \phi\alpha_1\alpha_2 + (1-\alpha_1)\alpha_3 - 1$ $\theta_{M+1} = \phi(1-\alpha_1) - \phi(1-\alpha_1)\alpha_3$	-
DM	-	-	-

2.4.2 MSOE State Space Model

Harrison (1967) gave the MSOE state space models that underpin SES and Holt's method. Harvey (1984) reached the same results from a different viewpoint. Following their steps, we identify, for the first time, an MSOE state space model that underpins the damped Holt's methods. No underlying MSOE state space models have been found for other methods in Table 2.1 although the basic structural model gave by Harvey (1984) leads to something fairly close to the additive Holt-Winters' method.

• **SES**

If the slowly changing level in equation (2.2) is assumed to follow a random walk process, the resulting model is in the MSOE state space form

$$Y_t = \mu_t + \epsilon_t, \tag{2.52a}$$

$$\mu_t = \mu_{t-1} + \eta_t, \tag{2.52b}$$

where ϵ_t and η_t are mutually independent white noise processes with zero mean and $E[\epsilon_t^2] = \sigma^2$ and $E[\eta_t^2] = \sigma_\eta^2$. Harrison (1967) proved that, for model (2.52), SES provides optimal forecasts and the optimal value of the smoothing parameter α is given by

$$\alpha = \frac{\sqrt{r^2 + 4r} - r}{2} \tag{2.53}$$

where $r = \sigma_\eta^2/\sigma^2$ is often referred to as the "signal to noise" ratio. As long as $0 < r < \infty$, α falls into the interval (0,1). When $\sigma^2 = 0$ (i.e., $r = \infty$), the optimal α is 1.

Harvey (1984) reached the same conclusion by showing that SES is equivalent to the steady-state KF for model (2.52). KF is said to be in a steady state if the MSE matrix $\mathbf{P}_t$ becomes constant (i.e., $\mathbf{P}_t = \mathbf{P}_{t-1} = \mathbf{P}$). Combining equations (2.34) and

(2.37) together, the steady-state KF for model (2.52) is given by

$$l_t = l_{t-1} + \frac{P + \sigma_\eta^2}{P + \sigma_\eta^2 + \sigma^2} \cdot (Y_t - l_{t-1}), \tag{2.54a}$$

$$P = \frac{P + \sigma_\eta^2}{P + \sigma_\eta^2 + \sigma^2} \cdot \sigma^2. \tag{2.54b}$$

P is used in place of the boldface $\mathbf{P}$ as now P is a scalar. Solving equation (2.54b) for P gives

$$P = \frac{\sqrt{r^2 + 4r} - r}{2} \cdot \sigma^2 = \alpha\sigma^2. \tag{2.55}$$

Substituting P into equation (2.54a) yields

$$l_t = l_{t-1} + \alpha(Y_t - l_{t-1}) = \alpha Y_t + (1 - \alpha) l_{t-1}, \tag{2.56}$$

which is the same as the recurrence equation for SES.

The optimality of SES for model (2.52) can also be seen from the fact that differencing Y_t once reduces model (2.52) to an ARIMA(0,1,1) model with

$$\theta_1 = \frac{\sqrt{r^2 + 4r} - r}{2} - 1. \tag{2.57}$$

With $r > 0$, we have $-1 < \theta_1 \leq 0$ (Harvey 1990). That is, model (2.52) is the state space representation of the ARIMA(0,1,1) model (2.45).

• Holt's Method

Assuming that both the level and the slope in equation (2.9) change slowly with time according to random walk processes leads to the MSOE state space model

$$Y_t = \mu_t + \epsilon_t, \tag{2.58a}$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \tag{2.58b}$$

$$\beta_t = \beta_{t-1} + \zeta_t, \tag{2.58c}$$

where ϵ_t, η_t and ζ_t are mutually independent white noise processes with zero mean and $E[\epsilon_t^2] = \sigma^2$, $E[\eta_t^2] = \sigma_\eta^2$ and $E[\zeta_t^2] = \sigma_\zeta^2$. Harrison (1967) proved that Holt's

method provides optimal forecasts for model (2.58) with the optimal values of the smoothing parameters α_1 and α_2 given by

$$r_\eta = \frac{\alpha_1^2 + \alpha_1^2\alpha_2 - 2\alpha_1\alpha_2}{1-\alpha_1}, \tag{2.59a}$$

$$r_\zeta = \frac{\alpha_1^2\alpha_2^2}{1-\alpha_1}, \tag{2.59b}$$

where $r_\eta = \sigma_\eta^2/\sigma^2$ and $r_\zeta = \sigma_\zeta^2/\sigma^2$.

Similar to the SES case, the optimality of Holt's method for model (2.58) can also be obtained by showing that

a). Holt's method with α_1 and α_2 given by equations (2.59a) and (2.59b) is equivalent to the steady-state KF for model (2.58) (Harvey 1990);

b). Model (2.58) is the state space representation of a subclass of the ARIMA(0,2,2) model underlying Holt's method.

Differencing Y_t twice reduces the state space model (2.58) to an ARIMA(0,2,2) model with θ_1 and θ_2 given by

$$r_\eta = -\frac{\theta_1 + \theta_1\theta_2 + 4\theta_2}{\theta_2}, \tag{2.60a}$$

$$r_\zeta = \frac{(1+\theta_1+\theta_2)^2}{\theta_2}. \tag{2.60b}$$

For $r_\eta > 0$ and $r_\zeta > 0$, we have

$$-(1+\theta_2) < \theta_1 < -\frac{4\theta_2}{1+\theta_2} \quad \text{and} \quad 0 \le \theta_2 < 1. \tag{2.61}$$

The condition $0 \le \theta_2 < 1$ implies that

$$-\frac{4\theta_2}{1+\theta_2} - (-2\theta_2) = \frac{2\theta_2(\theta_2-1)}{1+\theta_2} \le 0 \tag{2.62}$$

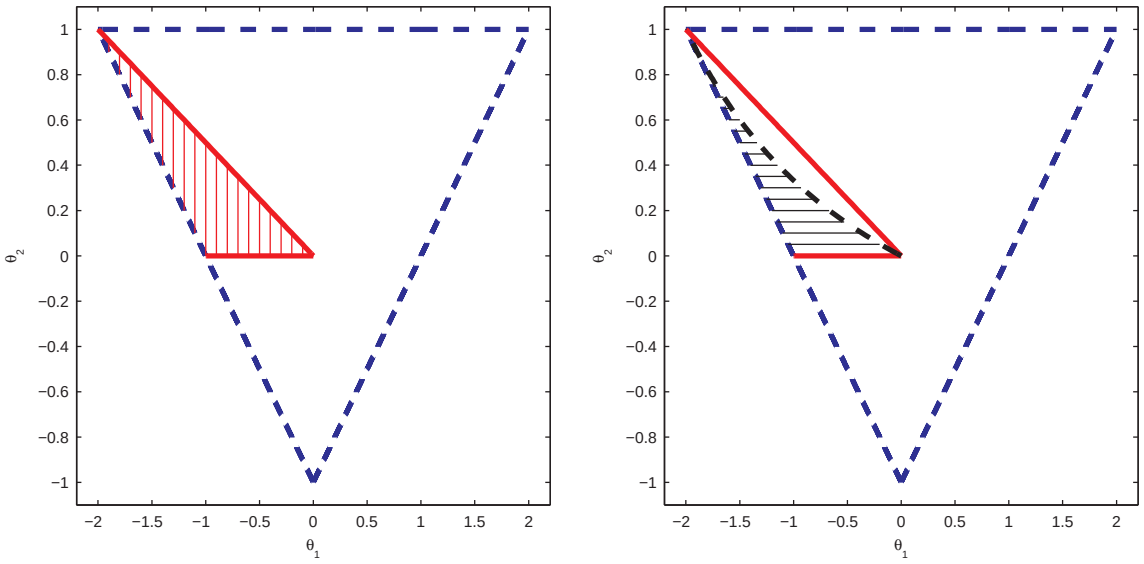
That is, the parameter space defined by (2.61) is a subset of the parameter space for the ARIMA(0,2,2) model (2.47) (see Figure 2.1).

As a special case of Holt's method, Brown's double exponential smoothing provides optimal forecasts for the state space model (2.58) when

$$r_\zeta = (r_\eta/2)^2, \tag{2.63}$$

and the optimal value of α is given by

$$\alpha = \frac{\sqrt{r_\eta^2 + 8r_\eta} - r_\eta}{4}. \tag{2.64}$$



(a) (b)

Figure 2.1: The big triangular area defines the parameter space for invertible ARIMA(0,2,2) model; the shaded area in (a) defines the parameter space for invertible ARIMA(0,2,2) model underlying Holt's method with α_1 and α_2 falling into the interval (0,1]; the shaded area in (b) defines the parameter space for the invertible ARIMA(0,2,2) model reduced from the state space model (2.58). (Dashed line – boundary not included; solid line – boundary included)

• Damped Holt's Method

Assuming that the level in equation (2.9) follows a random walk process while the slope follows an AR(1) process leads to the MSOE state space model

$$Y_t = \mu_t + \epsilon_t, \tag{2.65a}$$

$$\mu_t = \mu_{t-1} + \phi\beta_{t-1} + \eta_t, \tag{2.65b}$$

$$\beta_t = \phi\beta_{t-1} + \zeta_t, \tag{2.65c}$$

where the AR parameter ϕ takes values in the interval $[0, 1]$. It can be proven that, for model (2.65), the damped Holt's method with α_1 and α_2 satisfying

$$r_\eta = \frac{\alpha_1^2 + \phi\alpha_1^2\alpha_2 - (1+\phi)\alpha_1\alpha_2}{1-\alpha_1} \tag{2.66a}$$

$$r_\zeta = \frac{\phi^2\alpha_1^2\alpha_2^2 + \phi(1-\phi)\alpha_1^2\alpha_2 - (1-\phi)(1-\phi^2)\alpha_1\alpha_2}{\phi^2(1-\alpha_1)} \tag{2.66b}$$

provides optimal forecasts.

The optimality of damped Holt's method for model (2.65) can also be obtained by showing that

a). Damped Holt's method with α_1 and α_2 given by equations (2.66a) and (2.66b) is equivalent to the steady-state KF for model (2.65);

b). Model (2.65) is the state space representation of a subclass of the ARIMA(1,1,2) model underlying damped Holt's method.

2.4.3 SSOE State Space Model

Ord et al. (1997) introduced a very general form for SSOE state space models to encompass linear or nonlinear models with homoscedastic or heteroscedastic variance. With this general formulation, they explored the idea of using SSOE state space models as the underlying statistical models for ES methods, and identified for the first time a statistical model that underpins the multiplicative Holt-Winters' method. Since their work, a series of papers have been published on using SSOE models to explain ES methods. Chatfield et al. (2001) listed arguments in favor of SSOE models and discussed various possible SSOE models that underlie SES and the multiplicative Holt-Winters' method. Hyndman et al. (2002) provided two SSOE models for each method except these in the last row of Table 2.1, one model with homoscedastic variance and one model with heteroscedastic variance. We expand their work by adding underlying SSOE models for these methods in the last row of Table 2.1, namely the damped methods for time series with multiplicative trends.

SSOE state space models underlying ES methods can be written in a general form

$$Y_t = f(\boldsymbol{\beta}_{t-1}) + \epsilon_t, \tag{2.67a}$$

$$\boldsymbol{\beta}_t = g(\boldsymbol{\beta}_{t-1}) + w(\boldsymbol{\alpha}, \boldsymbol{\beta}_{t-1})\epsilon_t, \tag{2.67b}$$

where ϵ_t is a white noise process with $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = \sigma^2$, $\boldsymbol{\beta}_t$ is a $p \times 1$ state vector, $\boldsymbol{\alpha}$ is a vector of smoothing parameters, f is a mapping from $\Re^p$ to $\Re$, and g and w are mappings from $\Re^p$ to $\Re^p$. In addition, ϵ_t is assumed to be independent of $\boldsymbol{\beta}_k$ for any $k < t$. That is,

$$E[\epsilon_t \boldsymbol{\beta}_k] = 0, \text{ for } k < t. \tag{2.68}$$

Using SES as an example, we have $\boldsymbol{\beta}_t = \mu_t$, $f(\mu_t) = \mu_t$, $g(\mu_t) = \mu_t$, and $w(\boldsymbol{\alpha}, \mu_t) = \alpha$. As a result, the SSOE state space model in equations (2.67a) and (2.67b) becomes

$$Y_t = \mu_{t-1} + \epsilon_t, \tag{2.69a}$$

$$\mu_t = \mu_{t-1} + \alpha\epsilon_t. \tag{2.69b}$$

At time $t-1$, given μ_{t-1}, equation (2.69a) gives a one-step-ahead forecast

$$\hat{Y}_{t|t-1} = \mu_{t-1}. \tag{2.70}$$

Expressed in terms of the one-step-ahead forecast error $\epsilon_t = Y_t - \hat{Y}_{t|t-1}$, equation (2.69b), becomes

$$\mu_t = \mu_{t-1} + \alpha\epsilon_t. \tag{2.71}$$

Equations (2.70) and (2.71) are exactly the one-step-ahead forecast equation and the error-correction form updating equation of SES, albeit l_t and e_t replaced by their true values μ_t and ϵ_t respectively.

Table 2.4 contains underlying SSOE state space models for all of the ES methods listed in Table 2.1. Comparing Table 2.4 to Table 2.1 reveals that, for an ES method, the construction of an underlying SSOE state space model is straightforward, simply using the updating equation and the one-step-ahead forecasting equation with an

extra error term adding to its right side as the transition equation and the observation equation of the corresponding SSOE state space model respectively, albeit l_t, b_t, c_t, and e_t replaced by their true values μ_t, β_t, s_t, and ϵ_t (see Table 2.5).

Replacing ϵ_t in model (2.67) with $u(\boldsymbol{\beta}_{t-1})\epsilon_t$, where u is a mapping from $\Re^p$ to $\Re$, we end up with a more general class of SSOE state space models, which encompasses both homoscedastic (i.e., $u(\boldsymbol{\beta}_{t-1})$ is constant) and heteroscedastic (i.e., $u(\boldsymbol{\beta}_{t-1})$ is time varying) cases. Furthermore, for any mapping u, model (2.67) produces the same forecast. That is, the SSOE state space models that underpin an ES methods are not unique. This fact helps to explain the robustness of ES methods.

2.5 Discussion

Three types of statistical models have been found to underpin ES methods: ARIMA model, MSOE state space model, and SSOE state space model. Underlying ARIMA models exist for only additive ES methods, and often have smaller parameter spaces than unrestricted stationary and invertible ARIMA models do. Underlying MSOE state space models have been identified for SES, Holt's method, and the damped Holt's method. Moreover, as shown above, linear MSOE state space models can be reduced to an ARIMA form, and the reduced form ARIMA models are a subclass of these underlying ARIMA models for the corresponding ES methods (see Figure 2.1). However, MSOE state space models have the advantage of being able to incorporate explicitly structural features, which otherwise are hidden in ARIMA models by differencing. Keeping the structural form of MSOE models, SSOE state space models replace the independence assumption by a perfect correlation between the transient error and the permanent error. The perfect correlation assumption gives SSOE state space models advantages over ARIMA models and MSOE state space models. First, underlying SSOE state space models can be identified for all ES method in Table 2.1, and they are not unique and could be homoscedastic or heteroscedastic. Second,

Table 2.4: Underlying SSOE State Space Models for ES methods. $\xi_t = u(\boldsymbol{\beta}_{t-1})\epsilon_t$, and constant $u(\boldsymbol{\beta}_{t-1})$ gives homoscedastic models while time-varying $u(\boldsymbol{\beta}_{t-1})$ results in heteroscedastic models. (**N** - None, **A** - Additive, **M** - Multiplicative, **DA** - Damped Additive, **DM** - Damped Multiplicative)

	Seasonality		
Trend	**N**	**A**	**M**
N	$Y_t = \mu_{t-1} + \xi_t$	$Y_t = \mu_{t-1} + s_{t-M} + \xi_t$	$Y_t = \mu_{t-1}s_{t-M} + \xi_t$
	$\mu_t = \mu_{t-1} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1} + \alpha_1\xi_t/s_{t-M}$
		$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t$	$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t/\mu_{t-1}$
A	$Y_t = \mu_{t-1} + \beta_{t-1} + \xi_t$	$Y_t = \mu_{t-1} + \beta_{t-1} + s_{t-M} + \xi_t$	$Y_t = (\mu_{t-1} + \beta_{t-1})s_{t-M} + \xi_t$
	$\mu_t = \mu_{t-1} + \beta_{t-1} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1} + \beta_{t-1} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1} + \beta_{t-1} + \alpha_1\xi_t/s_{t-M}$
	$\beta_t = \beta_{t-1} + \alpha_1\alpha_2\xi_t$	$\beta_t = \beta_{t-1} + \alpha_1\alpha_2\xi_t$	$\beta_t = \beta_{t-1} + \alpha_1\alpha_2\xi_t/s_{t-M}$
		$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t$	$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t/(\mu_{t-1} + \beta_{t-1})$
M	$Y_t = \mu_{t-1}\beta_{t-1} + \xi_t$	$Y_t = \mu_{t-1}\beta_{t-1} + s_{t-M} + \xi_t$	$Y_t = (\mu_{t-1}\beta_{t-1})s_{t-M} + \xi_t$
	$\mu_t = \mu_{t-1}\beta_{t-1} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1}\beta_{t-1} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1}\beta_{t-1} + \alpha_1\xi_t/s_{t-M}$
	$\beta_t = \beta_{t-1} + \alpha_1\alpha_2\xi_t/\mu_{t-1}$	$\beta_t = \beta_{t-1} + \alpha_1\alpha_2\xi_t/\mu_{t-1}$	$\beta_t = \beta_{t-1} + \alpha_1\alpha_2\xi_t/(\mu_{t-1}s_{t-M})$
		$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t$	$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t/(\mu_{t-1}\beta_{t-1})$
DA	$Y_t = \mu_{t-1} + \phi\beta_{t-1} + \xi_t$	$Y_t = \mu_{t-1} + \phi\beta_{t-1} + s_{t-M} + \xi_t$	$Y_t = (\mu_{t-1} + \phi\beta_{t-1})s_{t-M} + \xi_t$
	$\mu_t = \mu_{t-1} + \phi\beta_{t-1} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1} + \phi\beta_{t-1} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1} + \phi\beta_{t-1} + \alpha_1\xi_t/s_{t-M}$
	$\beta_t = \phi\beta_{t-1} + \alpha_1\alpha_2\xi_t$	$\beta_t = \phi\beta_{t-1} + \alpha_1\alpha_2\xi_t$	$\beta_t = \phi\beta_{t-1} + \alpha_1\alpha_2\xi_t/s_{t-M}$
		$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t$	$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t/(\mu_{t-1} + \phi\beta_{t-1})$
DM	$Y_t = \mu_{t-1}\beta_{t-1}^{\phi} + \xi_t$	$Y_t = \mu_{t-1}\beta_{t-1}^{\phi} + s_{t-M} + \xi_t$	$Y_t = (\mu_{t-1}\beta_{t-1}^{\phi})s_{t-M} + \xi_t$
	$\mu_t = \mu_{t-1}\beta_{t-1}^{\phi} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1}\beta_{t-1}^{\phi} + \alpha_1\xi_t$	$\mu_t = \mu_{t-1}\beta_{t-1}^{\phi} + \alpha_1\xi_t/s_{t-M}$
	$\beta_t = \beta_{t-1}^{\phi} + \alpha_1\alpha_2\xi_t/\mu_{t-1}$	$\beta_t = \beta_{t-1}^{\phi} + \alpha_1\alpha_2\xi_t/\mu_{t-1}$	$\beta_t = \beta_{t-1}^{\phi} + \alpha_1\alpha_2\xi_t/(\mu_{t-1}s_{t-M})$
		$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t$	$s_t = s_{t-M} + (1-\alpha_1)\alpha_3\xi_t/(\mu_{t-1}\beta_{t-1}^{\phi})$

Table 2.5: Construction of Underlying SSOE State Space Models for ES methods ($\hat{\boldsymbol{\beta}}_t = (l_t, b_t, c_t, \cdots, c_{t-M+1})^T$, and $\boldsymbol{\beta}_t = (\mu_t, \beta_t, s_t, \cdots, s_{t-M+1})^T$)

ES method	**Underlying SSOE State Space Model**
Updating Equation	Transition Equation
$\hat{\boldsymbol{\beta}}_t = g(\hat{\boldsymbol{\beta}}_{t-1}) + w(\boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}_{t-1})e_t$	$\boldsymbol{\beta}_t = g(\boldsymbol{\beta}_{t-1}) + w(\boldsymbol{\alpha}, \boldsymbol{\beta}_{t-1})\xi_t$
One-Step-Ahead Forecast Equation	Observation Equation
$\hat{Y}_{t\|t-1} = f(\hat{\boldsymbol{\beta}}_{t-1})$	$Y_t = f(\boldsymbol{\beta}_{t-1}) + \epsilon_t$

the transition equations of underlying SSOE models provide a transparent link with the updating equations of the corresponding ES methods. Last, the class of underlying SSOE models, as shown in Table 2.6, is broader than the class of underlying ARIMA model, and therefore the class of underlying MSOE state space models. Snyder (1985) showed that any ARIMA model can be written in an SSOE state space form. SSOE state space models provides a formal statistical framework for the study of ES methods.

Table 2.6: Relationships among Three Types of Underlying Statistical Models for ES methods

class of underlying SSOE state space models
$\bigcup$
class of underlying ARIMA models
$\bigcup$
class of underlying MSOE state space models

CHAPTER III

PERFORMANCE ANALYSIS OF ES METHODS FOR TIME SERIES OF ARIMA TYPE

3.1 Introduction

Although the widespread usage of ES methods in industry and business for forecasting can be attributed to simple and intuitive formulation, efficient computation, and flexible adaptivity, a more important explanation of their popularity is that ES methods perform reasonably well for a wide class of time series. Comparing the performance of various forecasting methods on 111 real-world time series , Makridakis and Hibon (1979), found that statistically sophisticated methods do not necessarily provide more accurate forecasts than simple methods such as ES methods. Such a finding was not well received by statisticians. In response, Makridakis et al. launched two larger-scale empirical studies, M-Competition (Makridakis et al. 1982) and M2-Competition (Makridakis et al. 1993), by including more forecasting methods and larger number of time series. Both competitions once again concluded that, as simple as they are, ES methods such as SES and Holt's method perform as well as, or in many cases better than, statistically sophisticated methods such as the ARIMA models advocated by Box and Jenkins (Box et al. 1994) in terms of forecast accuracy. This conclusion was further confirmed by their latest empirical study, M3-Competition (Makridakis and Hibon 2000), which includes 3,003 time series of various types and 24 forecasting methods.

To better understand ES methods and be able to answer questions, such as when and why ES methods perform well, and how the choices of smoothing parameters

affect the performance, we investigate the behaviors of ES methods when time series are generated from different processes. We mainly focus on time series of ARIMA type. The purpose is to find out for what kinds of ARIMA-type time series ES methods perform well and why. Also studied are the effects of the values of the smoothing parameters on performance. For example, how the choices of the smoothing parameters affect the performance, when the values of smoothing parameters are critical for forecasting and when are not, and what will occur if the smoothing parameters are underestimated or overestimated.

Only included in this investigation are SES and Holt's linear trend method as these two are the most frequently used ones and were singled out by the M-competitions for their good performance. Also only one-step-ahead forecast is concerned for the performance analysis.

3.2 SES

SES updates the one-step-ahead forecast, $\hat{Y}_{t|t-1}$, by

$$\hat{Y}_{t|t-1} = \alpha Y_{t-1} + (1-\alpha)\hat{Y}_{t-1|t-2}, \tag{3.1}$$

where α is a smoothing parameter taking values in the interval (0,1].

Assume that the true data generating process follows a model of form

$$Y_t = \mu + N_t, \tag{3.2}$$

where μ is a constant, and N_t is a disturbance term with zero mean and first difference as a stationary process

$$(1-B)N_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}, \tag{3.3}$$

where B is the back shift operator, and ϵ_t is a zero-mean white noise process (i.e., $E[\epsilon_t] = 0, E[\epsilon_t^2] = \sigma^2$, and $E[\epsilon_t \epsilon_{t-k}] = 0$ for $k \neq 0$).

SES with a smoothing parameter α for model (3.2) leads to

$$e_t - (1-\alpha)e_{t-1} = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}. \tag{3.4}$$

Let $\lambda = 1 - \alpha$. As $0 < \alpha \leq 1$, $0 \leq \lambda < 1$. Therefore,

$$e_t = (1 - \lambda B)^{-1} \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \lambda^j \psi_i \epsilon_{t-i-j} = \sum_{j=0}^{\infty} (\sum_{i=0}^{j} \lambda^{j-i} \psi_i) \epsilon_{t-j}. \tag{3.5}$$

As a result, the mean of the one-step-ahead forecast error is given by

$$E[e_t] = \sum_{j=0}^{\infty} (\sum_{i=0}^{j} \lambda^{j-i} \psi_i) E[\epsilon_{t-j}] = 0, \tag{3.6}$$

which implies that the one-step-ahead forecast by SES for model (3.2) is unbiased. The mean squared error (MSE) of the one-step-ahead forecast is given by

$$E[e_t^2] = \sigma^2 \sum_{j=0}^{\infty} (\sum_{i=0}^{j} \lambda^{j-i} \psi_i)^2. \tag{3.7}$$

3.2.1 N_t is an ARIMA$(0, 1, q)$ process

N_t is an ARIMA$(0, 1, q)$ process

$$(1 - B) N_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i} \tag{3.8}$$

Without loss of generality, we assume that $\theta_0 = 1$. Therefore, $\psi_i = \theta_i$ for $0 \leq i \leq q$, and $\psi_i = 0$ for $i > q$. The MSE in equation (3.7) becomes

$$E[e_t^2] = \sigma^2 [\sum_{j=0}^{q-1} (\sum_{i=0}^{j} \lambda^{j-i} \theta_i)^2 + \frac{1}{1 - \lambda^2} (\sum_{i=0}^{q} \lambda^{q-i} \theta_i)^2] \tag{3.9}$$

1). $q = 0$, N_t is an ARIMA(0,1,0), which is a random walk. That is, $N_t = N_{t-1} + \epsilon_t$.

$$E[e_t^2] = \sigma^2 \cdot \frac{1}{1 - \lambda^2}. \tag{3.10}$$

According to equation (3.10) as well as Figure 3.1, MSE is a monotone increasing function of λ for $\lambda \in [0, 1)$, therefore a monotone decreasing function of α for $\alpha \in (0, 1]$. The minimum MSE occurs as $\alpha = 1 (\lambda = 0)$, at which the one-step-ahead forecast is $\hat{Y}_{t|t-1} = Y_{t-1}$. That is, the forecast for the value at time t is simply the observed value at time $t - 1$. This is due to the property of the random walk

$$E[Y_t | Y_{t-1}, Y_{t-2}, \cdots] = E[Y_t | Y_{t-1}], \tag{3.11}$$

which implies that, given the value at time $t-1$, the value at time t is independent of the values at times $t-k, k>1$.

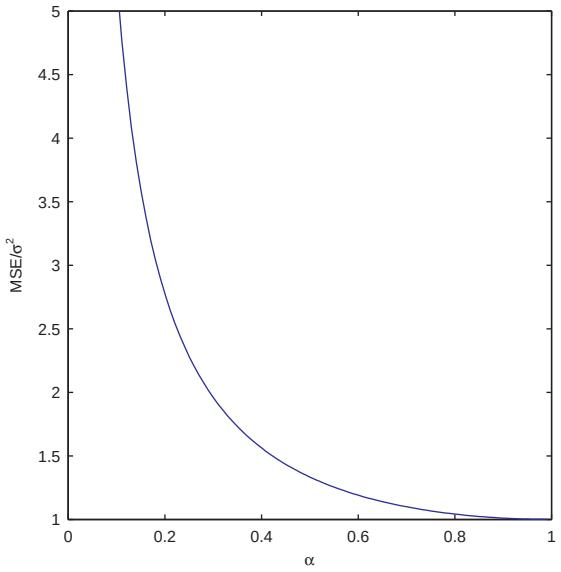


Figure 3.1: SES – MSE/σ^2 as a function of α, N_t is an ARIMA(0,1,0).

2). $q=1$, N_t is an ARIMA(0,1,1).

$$E[e_t^2] = \sigma^2 \cdot [1 + \frac{(\lambda+\theta_1)^2}{1-\lambda^2}]. \tag{3.12}$$

Figure 3.2 and Table 3.1 show that

- SES performs better when $-1 < \theta_1 < 0$ than when $0 < \theta_1 < 1$. When $-1 < \theta_1 < 0$, the minimum MSE stays at the theoretically optimal value σ^2. When $0 < \theta_1 < 1$, the minimum MSE is greater than σ^2 and increases as θ_1 increases.
- When $-1 < \theta_1 < 0$, the minimum MSE occurs as $\alpha = 1+\theta_1$ (i.e., $\lambda = -\theta_1$). When $0 < \theta_1 < 1$, MSE is a monotone decreasing function of α and reaches minimum at $\alpha = 1$.
- Overestimation of α is less serious than the equivalent underestimation.

In fact, SES is optimal for an ARIMA(0,1,1)

$$(1 - B)N_t = \epsilon_t + \theta_1 \epsilon_{t-1} \tag{3.13}$$

with

$$-1 < \theta_1 < 0 \tag{3.14}$$

when the smoothing parameter $\alpha = 1 + \theta_1$ (Muth 1960). Let α_{opt} denote the value of α, at which the minimum MSE is achieved. Taking the derivative of $E[e_t^2]$ in equation (3.12) with respect to λ and setting it to zero gives

$$\alpha_{opt} = \begin{cases} 1 + \theta_1, & -1 < \theta_1 < 0, \\ 1, & 0 < \theta_1 < 1. \end{cases} \tag{3.15}$$

The MSE at α_{opt} (i.e., the minimum MSE) is

$$\text{MSE}_{min} = \begin{cases} \sigma^2, & -1 < \theta_1 < 0, \\ \sigma^2(1 + \theta_1^2), & 0 < \theta_1 < 1. \end{cases} \tag{3.16}$$

In summary, SES is optimal for an ARIMA(0,1,1) when $-1 < \theta_1 < 0$ and performs well when $0 < \theta_1 < 1$ but has a small value. Also, a large α seems a safe choice.

Table 3.1: SES – Optimal α and Minimum MSE/σ^2, N_t is an ARIMA(0,1,1)

θ_1	α_{opt}	MSE/σ^2	θ_1	α_{opt}	MSE/σ^2
-0.95	0.05	1.00	0.95	1.00	1.90
-0.90	0.10	1.00	0.90	1.00	1.81
-0.80	0.20	1.00	0.80	1.00	1.64
-0.70	0.30	1.00	0.70	1.00	1.49
-0.60	0.40	1.00	0.60	1.00	1.36
-0.50	0.50	1.00	0.50	1.00	1.25
-0.40	0.60	1.00	0.40	1.00	1.16
-0.30	0.70	1.00	0.30	1.00	1.09
-0.20	0.80	1.00	0.20	1.00	1.04
-0.10	0.90	1.00	0.10	1.00	1.01

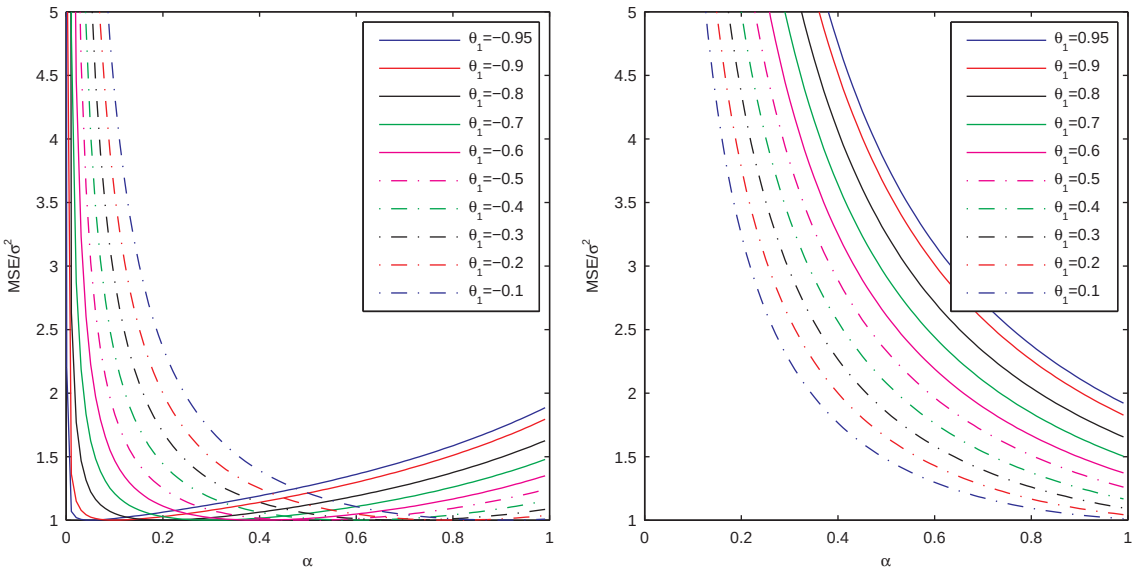


Figure 3.2: SES – MSE/σ^2 as a function of α, N_t is an ARIMA(0,1,1)

3). $q = 2$, N_t is an ARIMA(0,1,2).

$$E[e_t^2] = \sigma^2 \cdot [1 + (\lambda + \theta_1)^2 + \frac{(\lambda^2 + \lambda\theta_1 + \theta_2)^2}{1 - \lambda^2}]. \tag{3.17}$$

From Figure 3.3 and Table 3.2, the following conclusions can be drawn:

- Given θ_1, the minimum MSE increases as $|\theta_2|$, the absolute value of θ_2, increases. Given θ_2, the minimum MSE remains unchanged when $\theta_1 \leq 0$ and increases with θ_1 when $\theta_1 > 0$.
- When $\theta_1 < 0$, α_{opt} is less than 1 and increases as θ_1 and/or θ_2 increase. When $\theta_1 \geq 0$, $\alpha_{opt} = 1$.
- Overestimation of α is less serious than the equivalent underestimation.

Taking the derivative of $E[e_t^2]$ in equation (3.17) with respect to λ and setting the result to zero gives

$$\alpha_{opt} = \begin{cases} (1 + \theta_1 + \theta_2)/(1 + \theta_2), & -(1 + \theta_2) < \theta_1 < 0, \\ 1, & 0 \leq \theta_1 < 1 + \theta_2. \end{cases} \tag{3.18}$$

The minimum MSE is

$$\text{MSE}_{min} = \begin{cases} \sigma^2(1+\theta_2^2), & -(1+\theta_2) < \theta_1 < 0, \\ \sigma^2(1+\theta_1^2+\theta_2^2), & 0 \le \theta_1 < 1+\theta_2. \end{cases} \tag{3.19}$$

In summary, SES performs well for an ARIMA(0,1,2) when θ_1 is negative and $|\theta_2|$ is small. A large α is a safe choice.

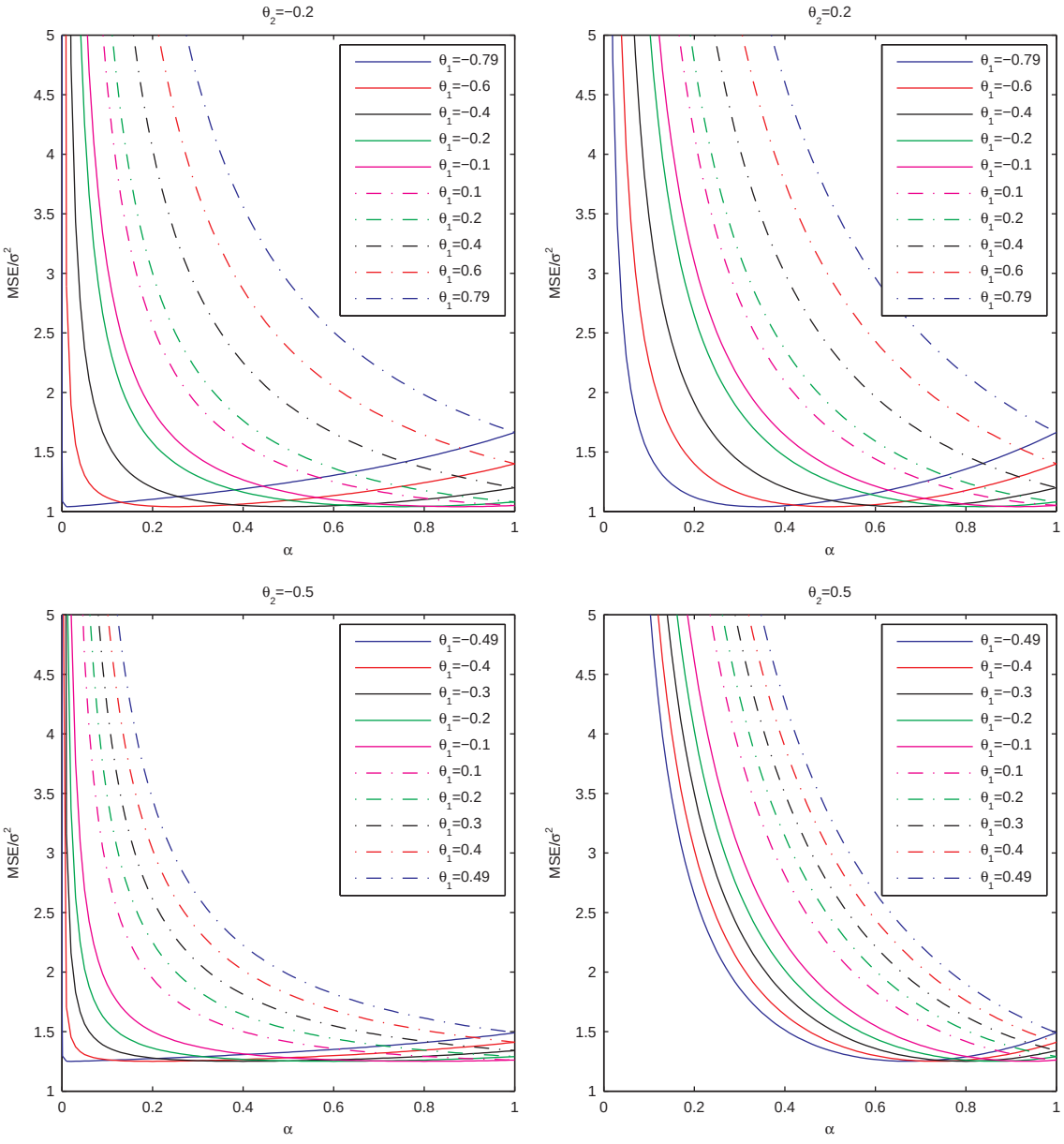


Figure 3.3: SES – MSE/σ^2 as a function of α, N_t is an ARIMA(0,1,2)

Table 3.2: SES – Optimal α and Minimum MSE/σ^2, N_t is an ARIMA(0,1,2)

	$\theta_2 = -0.2$		$\theta_2 = 0.2$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
-1.19	-	-	0.01	1.04
-1.00	-	-	0.17	1.04
-0.79	0.01	1.04	0.34	1.04
-0.60	0.25	1.04	0.50	1.04
-0.49	0.39	1.04	0.59	1.04
-0.40	0.50	1.04	0.67	1.04
-0.20	0.75	1.04	0.83	1.04
0.00	1.00	1.04	1.00	1.04
0.20	1.00	1.08	1.00	1.08
0.40	1.00	1.20	1.00	1.20
0.49	1.00	1.28	1.00	1.28
0.60	1.00	1.40	1.00	1.40
0.79	1.00	1.66	1.00	1.66
1.00	-	-	1.00	2.04
1.19	-	-	1.00	2.46

	$\theta_2 = -0.5$		$\theta_2 = 0.5$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
-1.49	-	-	0.01	1.25
-1.40	-	-	0.07	1.25
-1.19	-	-	0.21	1.25
-1.00	-	-	0.33	1.25
-0.79	-	-	0.47	1.25
-0.60	-	-	0.60	1.25
-0.49	0.02	1.25	0.67	1.25
-0.40	0.20	1.25	0.73	1.25
-0.20	0.60	1.25	0.87	1.25
0.00	1.00	1.25	1.00	1.25
0.20	1.00	1.29	1.00	1.29
0.40	1.00	1.41	1.00	1.41
0.49	1.00	1.49	1.00	1.49
0.60	-	-	1.00	1.61
0.79	-	-	1.00	1.87
1.00	-	-	1.00	2.25
1.19	-	-	1.00	2.67
1.40	-	-	1.00	3.21
1.49	-	-	1.00	3.47

3.2.2 N_t is an ARIMA$(0,0,q)$ process

N_t is an ARIMA$(0,0,q)$ process

$$N_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i}. \tag{3.20}$$

Let $\theta_{-1} = \theta_{q+1} = 0$, we have

$$(1-B)N_t = \sum_{i=0}^{q+1} (\theta_i - \theta_{i-1})\epsilon_{t-i}. \tag{3.21}$$

The MSE becomes

$$E[e_t^2] = \sigma^2[\sum_{j=0}^{q}(\sum_{i=0}^{j} \lambda^{j-i}(\theta_i - \theta_{i-1}))^2 + \frac{1}{1-\lambda^2}(\sum_{i=0}^{q+1} \lambda^{q+1-i}(\theta_i - \theta_{i-1}))^2]. \tag{3.22}$$

1). $q = 0$, N_t is a white noise.

$$E[e_t^2] = \sigma^2 \cdot \frac{2}{1+\lambda}. \tag{3.23}$$

According to equation 3.23 as well as Figure 3.4, MSE is a monotone decreasing function of λ, therefore a monotone increasing function of α. MSE $\to \sigma^2$ as $\alpha \to 0$ ($\lambda \to 1$). The same result was obtained by Cohen (1963). SES with $\alpha \to 0$ is analogous to taking the average of a random sample with sample size $n \to \infty$.

In fact, with a white noise disturbance, model (3.2) becomes

$$Y_t = \mu + \epsilon_t, \tag{3.24}$$

for which the average of a random sample gives the minimum variance estimator of $E[Y_t] = \mu$.

2). $q = 1$, N_t is a MA(1).

$$E[e_t^2] = \sigma^2 \cdot [2\theta_1 + \frac{2(1-\theta_1)^2}{1+\lambda}]. \tag{3.25}$$

Figure 3.5 and Table 3.3 show that

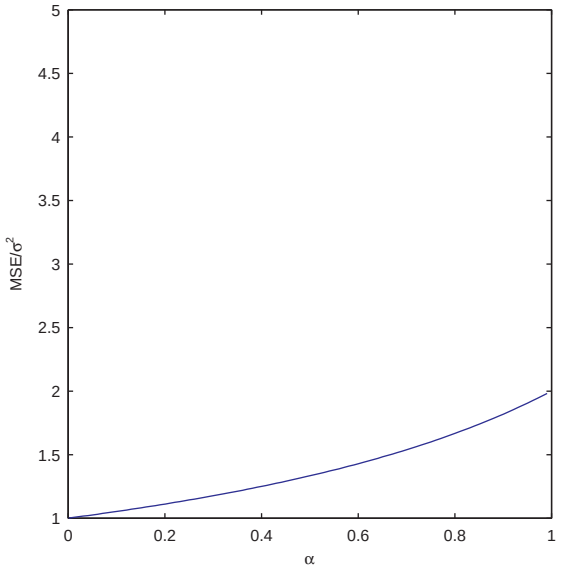


Figure 3.4: SES – MSE/σ^2 as a function of α, N_t is a white noise

- The MSE is a monotone increasing function of α and reaches minimum as $\alpha \to 0$ ($\lambda \to 1$).
- According to equation (3.25), as $\alpha \to 0$ ($\lambda \to 1$), MSE $\to \sigma^2(1+\theta_1^2)$. That is, the minimum MSE only depends on $|\theta_1|$, the absolute value of θ_1, and increases as $|\theta_1|$ increases.
- The larger the θ_1, the less critical the choice of α. For $\theta_1 > 0.5$, the MSE varies extremely slowly with α.

In summary, SES performs well for an MA(1) when $|\theta_1|$ is small. A small α is preferred.

3). $q = 2$, N_t is an MA(2).

$$E[e_t^2] = \sigma^2 \cdot [2(\theta_1 + \theta_1\theta_2 - 2\theta_2) + 2\theta_2\lambda + \frac{2(1-\theta_1+\theta_2)^2}{1+\lambda}]. \tag{3.26}$$

Figure 3.6 and Table 3.4 show that

- SES performs better when both $|\theta_1|$ and $|\theta_2|$ are small. Given θ_2 (or θ_1), the smaller the $|\theta_1|$ (or $|\theta_2|$), the smaller the minimum MSE.

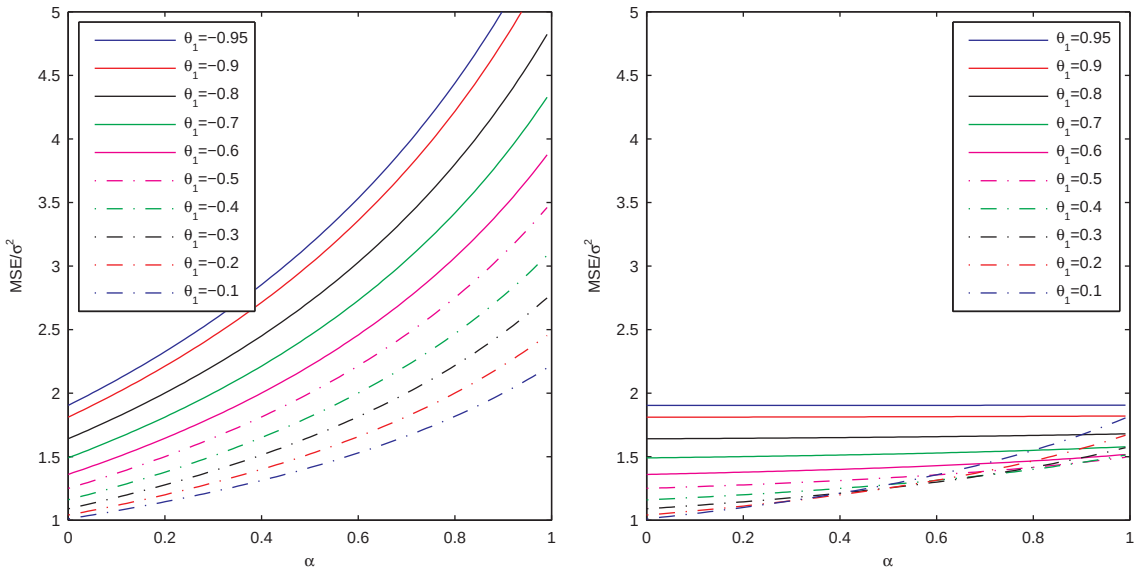


Figure 3.5: SES – MSE/σ^2 as a function of α, N_t is an MA(1)

Table 3.3: SES – Optimal α and Minimum MSE/σ^2, N_t is an MA(1)

θ_1	α_{opt}	MSE/σ^2	θ_1	α_{opt}	MSE/σ^2
-0.95	$\to 0$	1.90	0.95	$\to 0$	1.90
-0.90	$\to 0$	1.81	0.90	$\to 0$	1.81
-0.80	$\to 0$	1.64	0.80	$\to 0$	1.64
-0.70	$\to 0$	1.49	0.70	$\to 0$	1.49
-0.60	$\to 0$	1.36	0.60	$\to 0$	1.36
-0.50	$\to 0$	1.25	0.50	$\to 0$	1.25
-0.40	$\to 0$	1.16	0.40	$\to 0$	1.16
-0.30	$\to 0$	1.09	0.30	$\to 0$	1.09
-0.20	$\to 0$	1.04	0.20	$\to 0$	1.04
-0.10	$\to 0$	1.01	0.10	$\to 0$	1.01

– $\alpha_{opt} \to 0$ when θ_1 is small and $|\theta_2| < 1$.

– The choice of α is not critical when $\theta_1 > 0$ and $|\theta_2|$ is small.

It can be shown that, using Equation (3.26),

$$\alpha_{opt} = \begin{cases} \to 0, & -1 < \theta_2 \leq 0, \\ \to 0, & 0 < \theta_2 < 1,\ -(1+\theta_2) < \theta_1 \leq (1-\sqrt{\theta_2})^2, \\ \frac{\theta_1 - (1-\sqrt{\theta_2})^2}{\sqrt{\theta_2}}, & 0 < \theta_2 < 1,\ (1-\sqrt{\theta_2})^2 < \theta_1 \leq 1 + \theta_2 - \sqrt{\theta_2} \\ 1, & otherwise. \end{cases} \tag{3.27}$$

And, as $\alpha_{opt} \to 0$ ($\lambda \to 1$), the minimum MSE $\to \sigma^2(1 + \theta_1^2 + \theta_2^2)$.

In summary, SES performs well for an MA(2) only when both $|\theta_1|$ and $|\theta_2|$ are small, and a small α is a safe choice although the choice of α is not critical when both θ_1 and θ_2 are positive.

Table 3.4: SES – Optimal α and Minimum MSE/σ^2, N_t is a MA(2)

	$\theta_2 = -0.2$		$\theta_2 = 0.2$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
-1.19	-	-	$\to 0$	2.46
-1.00	-	-	$\to 0$	2.04
-0.79	$\to 0$	1.66	$\to 0$	1.66
-0.60	$\to 0$	1.40	$\to 0$	1.40
-0.49	$\to 0$	1.28	$\to 0$	1.28
-0.40	$\to 0$	1.20	$\to 0$	1.20
-0.20	$\to 0$	1.08	$\to 0$	1.08
0.00	$\to 0$	1.04	$\to 0$	1.04
0.20	$\to 0$	1.08	$\to 0$	1.08
0.40	$\to 0$	1.20	0.21	1.19
0.49	$\to 0$	1.28	0.41	1.25
0.60	$\to 0$	1.40	0.66	1.31
0.79	$\to 0$	1.66	1.00	1.43
1.00	-	-	1.00	1.68
1.19	-	-	1.00	2.06

	$\theta_2 = -0.5$		$\theta_2 = 0.5$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
-1.49	-	-	$\to 0$	3.47
-1.40	-	-	$\to 0$	3.21
-1.19	-	-	$\to 0$	2.67
-1.00	-	-	$\to 0$	2.25
-0.79	-	-	$\to 0$	1.87
-0.60	-	-	$\to 0$	1.61
-0.49	$\to 0$	1.49	$\to 0$	1.49
-0.40	$\to 0$	1.41	$\to 0$	1.41
-0.20	$\to 0$	1.29	$\to 0$	1.29
0.00	$\to 0$	1.25	$\to 0$	1.25
0.20	$\to 0$	1.29	0.16	1.28
0.40	$\to 0$	1.41	0.44	1.31
0.49	$\to 0$	1.49	0.57	1.33
0.60	-	-	0.73	1.35
0.79	-	-	1.00	1.38
1.00	-	-	1.00	1.50
1.19	-	-	1.00	1.76
1.40	-	-	1.00	2.22
1.49	-	-	1.00	2.47

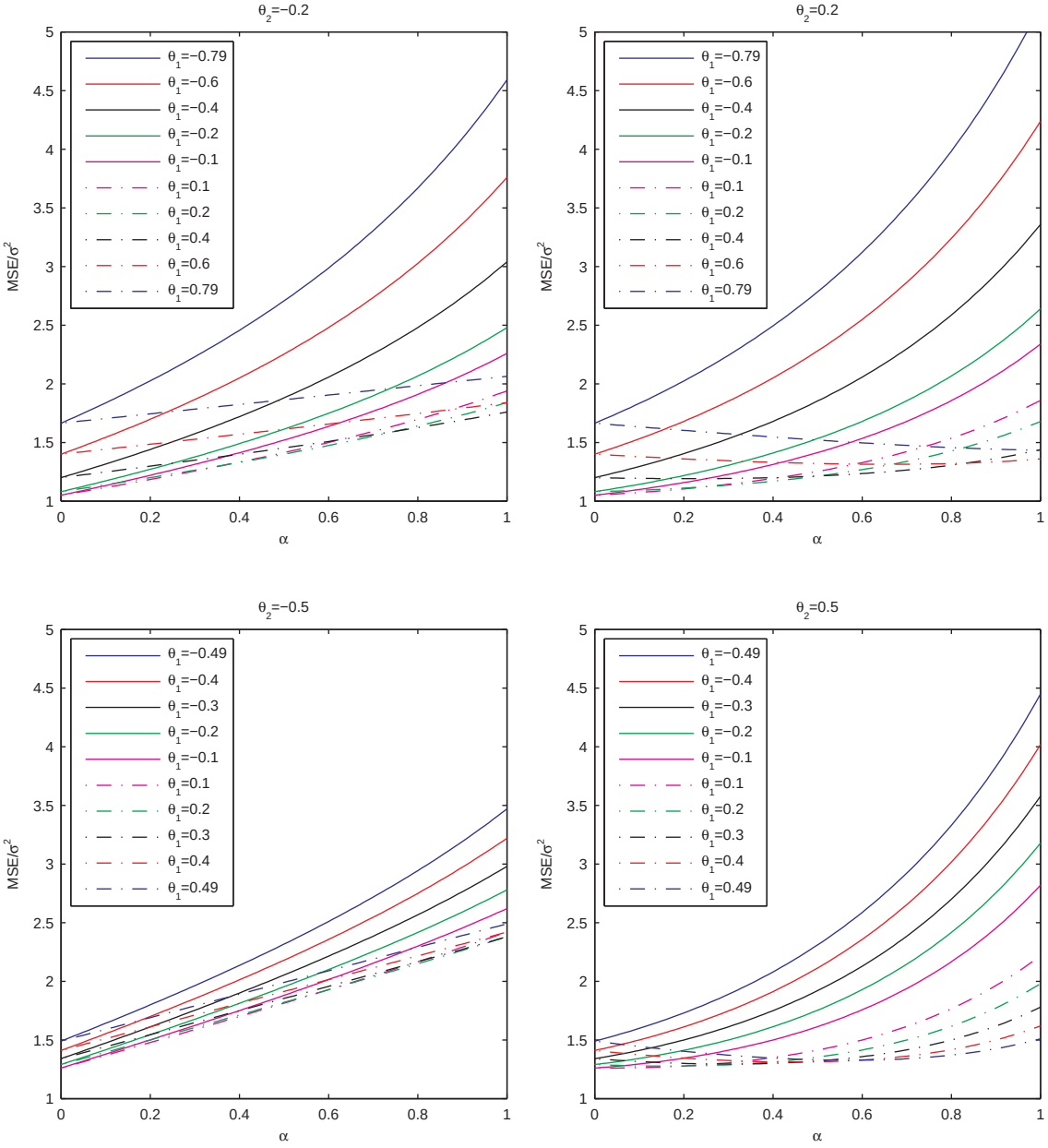


Figure 3.6: SES – MSE/σ^2 as a function of α, N_t is an MA(2)

3.2.3 N_t is an ARIMA$(1, d, 0)$ process

N_t is an ARIMA$(1, d, 0)$ process

$$(1 - B)^d N_t = (1 - \phi_1 B)^{-1} \epsilon_t. \tag{3.28}$$

1). $d = 1$, N_t is an ARIMA$(1, 1, 0)$.

$$(1 - B)N_t = (1 - \phi_1 B)^{-1} \epsilon_t = \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}. \tag{3.29}$$

Then, $\psi_i = \phi_1^i$ for $i \geq 0$. The MSE becomes

$$E[e_t^2] = \sigma^2 \sum_{j=0}^{\infty} (\sum_{i=0}^{j} \lambda^{j-i} \phi_1^i)^2 = \sigma^2 \cdot \frac{(1 + \lambda\phi_1)}{(1 - \lambda^2)(1 - \phi_1^2)(1 - \lambda\phi_1)}. \tag{3.30}$$

Figure 3.7 and Table 3.5 show that

- SES performs better when $-1 < \phi_1 < 0$ than when $0 < \phi_1 < 1$. When $-1 < \phi_1 < 0$, the minimum MSE increases as $|\phi_1|$, the absolute value of ϕ_1, increases, and the larger the $|\phi_1|$, the faster the minimum MSE grows. Same holds when $0 < \phi_1 < 1$.
- When $-0.6 < \phi_1 < 0$, the minimum MSE approximately equals to σ^2, and changes in ϕ_1 have little effect on the MSE. Moveover, a good choice of α is between 0.6 and 0.8.
- α_{opt} increases as ϕ_1 increases and reaches 1 at $\phi_1 = 0$.
- When $-1 < \phi_1 < 0$, overestimation of α is less serious than the equivalent underestimation especially when $\phi_1 \in (-0.6, 0)$. When $0 < \phi_1 < 1$, $\alpha_{opt} = 1$ and $\text{MSE}_{min} = \sigma^2/(1 - \phi_1^2)$.

In summary, SES performs well for an ARIMA(1,1,0) when $\phi_1 < 0$ and $|\phi_1|$ is small, and a large α is a safe choice.

2). $d = 0$, N_t is an AR(1).

$$N_t = (1 - \phi_1 B)^{-1} \epsilon_t, \tag{3.31}$$

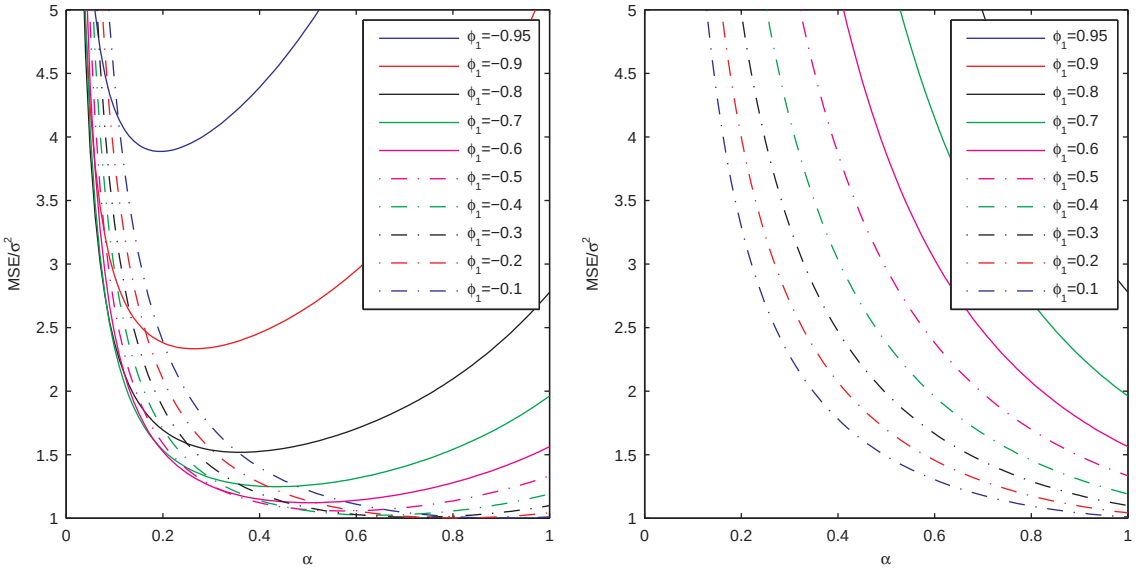


Figure 3.7: SES – MSE/σ^2 as a function of α, N_t is an ARIMA(1,1,0)

Table 3.5: SES – Optimal α and Minimum MSE/σ^2, N_t is an ARIMA(1,1,0)

ϕ_1	α_{opt}	MSE/σ^2
-0.95	0.19	3.88
-0.90	0.26	2.33
-0.80	0.36	1.52
-0.70	0.44	1.25
-0.60	0.50	1.12
-0.50	0.57	1.06
-0.40	0.64	1.02
-0.30	0.72	1.01
-0.20	0.81	1.00
-0.10	0.90	1.00

ϕ_1	α_{opt}	MSE/σ^2
0.95	1.00	10.26
0.90	1.00	5.26
0.80	1.00	2.78
0.70	1.00	1.96
0.60	1.00	1.56
0.50	1.00	1.33
0.40	1.00	1.19
0.30	1.00	1.10
0.20	1.00	1.04
0.10	1.00	1.01

which gives

$$(1-B)N_t = (1-B)(1-\phi_1 B)^{-1}\epsilon_t = \epsilon_t + \sum_{i=1}^{\infty} \phi_1^{i-1}(\phi_1 - 1)\epsilon_{t-i}. \tag{3.32}$$

Therefore, $\psi_0 = 1, \psi_i = \phi_1^{i-1}(\phi_1 - 1)$ for $i \geq 1$. The MSE becomes

$$E[e_t^2] = \sigma^2 \sum_{j=0}^{\infty} (\lambda^j + (\phi_1 - 1)\sum_{i=1}^{j} \lambda^{j-i}\phi_1^{i-1})^2 = \sigma^2 \cdot \frac{2}{(1+\lambda)(1+\phi_1)(1-\lambda\phi_1)}. \tag{3.33}$$

Figure 3.8 and Table 3.6 show that

- SES performs better when $0 < \phi_1 < 1$ than when $-1 < \phi_1 < 0$. When $-1 < \phi_1 < 0$, the smaller the ϕ_1, the larger the minimum MSE. When $0 < \phi_1 < 1$, the worse performance happens at $\phi_1 = 0.5$.
- When $-1 < \phi_1 < 0.4$, $\alpha_{opt} \to 0$. When $0.4 \leq \phi_1 < 1$, α_{opt} increases as ϕ_1 increase.

  In fact, Cohen (1963) showed that, for an AR(1) disturbance,

$$\alpha_{opt} = \begin{cases} (3\phi_1 - 1)/2\phi_1, & 1/3 < \phi_1 < 1, \\ \to 0, & -1 < \phi_1 \leq 1/3, \end{cases} \tag{3.34}$$

  and the corresponding minimum MSE is

$$E[e_t^2] = \begin{cases} 8\sigma^2\phi_1/(1+\phi_1)^3, & 1/3 < \phi_1 < 1, \\ \sigma^2/(1-\phi_1^2), & -1 < \phi_1 \leq 1/3. \end{cases} \tag{3.35}$$

  Cox (1961) conducted a rather detailed study on the performance of SES for an AR(1). However, the formula he gave for the MSE is not correct.
- The choice of α is not critical when ϕ_1 is close to 0.5. When $\phi_1 = 0.5$, the MSE is rather insensitive to the choice of α (see Table 3.7).

In summary, SES performs well for an AR(1) when $0 < \phi_1 < 1$, and the choice of α is not critical when ϕ_1 is close to 0.5.

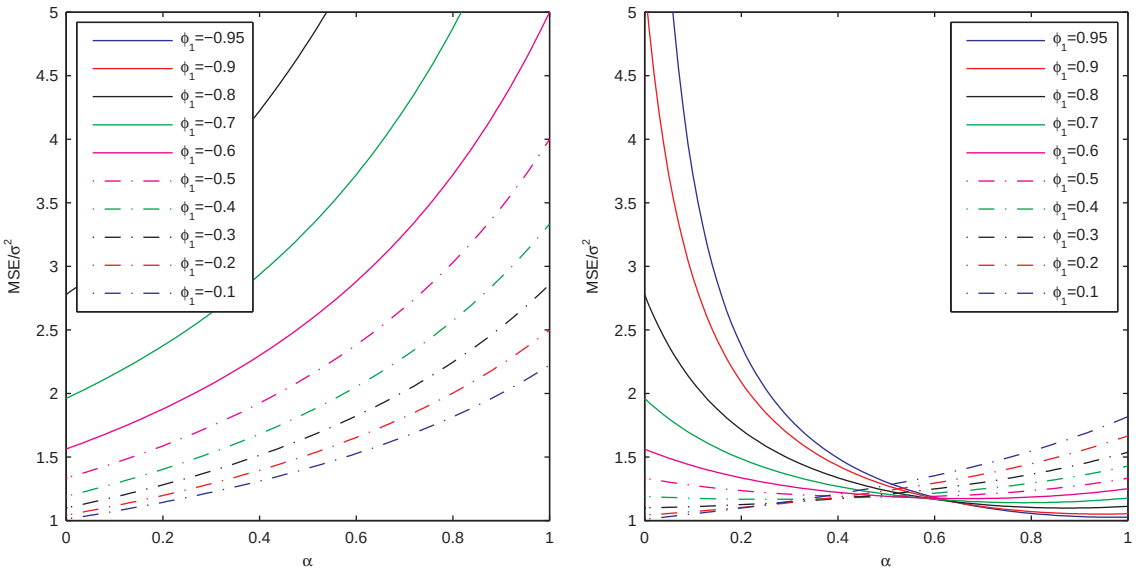


Figure 3.8: SES – MSE/σ^2 as a function of α, N_t is an AR(1)

Table 3.6: SES – Optimal α and Minimum MSE/σ^2, N_t is an AR(1)

ϕ_1	α_{opt}	MSE/σ^2	ϕ_1	α_{opt}	MSE/σ^2
-0.95	$\to 0$	10.26	0.95	0.97	1.02
-0.90	$\to 0$	5.26	0.90	0.94	1.05
-0.80	$\to 0$	2.78	0.80	0.88	1.10
-0.70	$\to 0$	1.96	0.70	0.79	1.14
-0.60	$\to 0$	1.56	0.60	0.67	1.17
-0.50	$\to 0$	1.33	0.50	0.50	1.19
-0.40	$\to 0$	1.19	0.40	0.25	1.17
-0.30	$\to 0$	1.10	0.30	$\to 0$	1.10
-0.20	$\to 0$	1.04	0.20	$\to 0$	1.04
-0.10	$\to 0$	1.01	0.10	$\to 0$	1.01

Table 3.7: SES – MSE/σ^2 for $\phi_1 = 0.5$, N_t is an AR(1), $\alpha_{opt} = 0.5$.

α	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
MSE/σ^2	1.30	1.28	1.23	1.21	1.19	1.19	1.19	1.21	1.23	1.28	1.30

3.2.4 N_t is an ARIMA$(1, d, 1)$ process

N_t is an ARIMA$(1, d, 1)$ process

$$(1 - B)^d N_t = (1 - \phi_1 B)^{-1}(\epsilon_t + \theta_1 \epsilon_{t-1}). \tag{3.36}$$

1). $d = 1$, N_t is an ARIMA(1,1,1).

$$(1 - B)N_t = (1 - \phi_1 B)^{-1}(\epsilon_t + \theta_1 \epsilon_{t-1}) = \epsilon_t + (\phi_1 + \theta_1) \sum_{i=1}^{\infty} \phi_1^{i-1} \epsilon_{t-i}. \tag{3.37}$$

Then, $\psi_0 = 1, \psi_i = (\phi_1 + \theta_1)\phi_1^{i-1}$ for $i \geq 1$. The MSE becomes

$$\begin{aligned} E[e_t^2] &= \sigma^2 \sum_{j=0}^{\infty} (\lambda^j + (\phi_1 + \theta_1) \sum_{i=1}^{j} \lambda^{j-i} \phi_1^{i-1})^2 \\ &= \sigma^2 \sum_{j=0}^{\infty} (\frac{\lambda^j(\lambda + \theta_1) - \phi_1^j(\phi_1 + \theta_1)}{\lambda - \phi_1})^2 \\ &= \sigma^2 \cdot \frac{(\lambda + \theta_1)(\phi_1 + \theta_1) + (1 + \lambda\theta_1)(1 + \phi_1\theta_1)}{(1 - \lambda^2)(1 - \phi_1^2)(1 - \lambda\phi_1)}. \end{aligned} \tag{3.38}$$

Figure 3.9 and Table 3.8 show that

- SES performs well under two situations: i) $-1 < \theta_1 < 0$ and $|\phi_1|$ is small, and ii) $0 < \theta_1 < 1$, $-1 < \phi_1 < 0$, and $|\theta_1|$ and $|\phi_1|$ are close.
- α_{opt} increases as both θ_1 and ϕ_1 increase, and overestimation of α is less serious than the equivalent underestimation.

2). $d = 0$, N_t is an ARMA(1,1).

$$N_t = (1 - \phi_1 B)^{-1}(\epsilon_t + \theta_1 \epsilon_{t-1}), \tag{3.39}$$

which gives

$$(1 - B)N_t = \epsilon_t + (\phi_1 + \theta_1 - 1)\epsilon_{t-1} + (\phi_1 + \theta_1)(\phi_1 - 1) \sum_{i=2}^{\infty} \phi_1^{i-2} \epsilon_{t-i}. \tag{3.40}$$

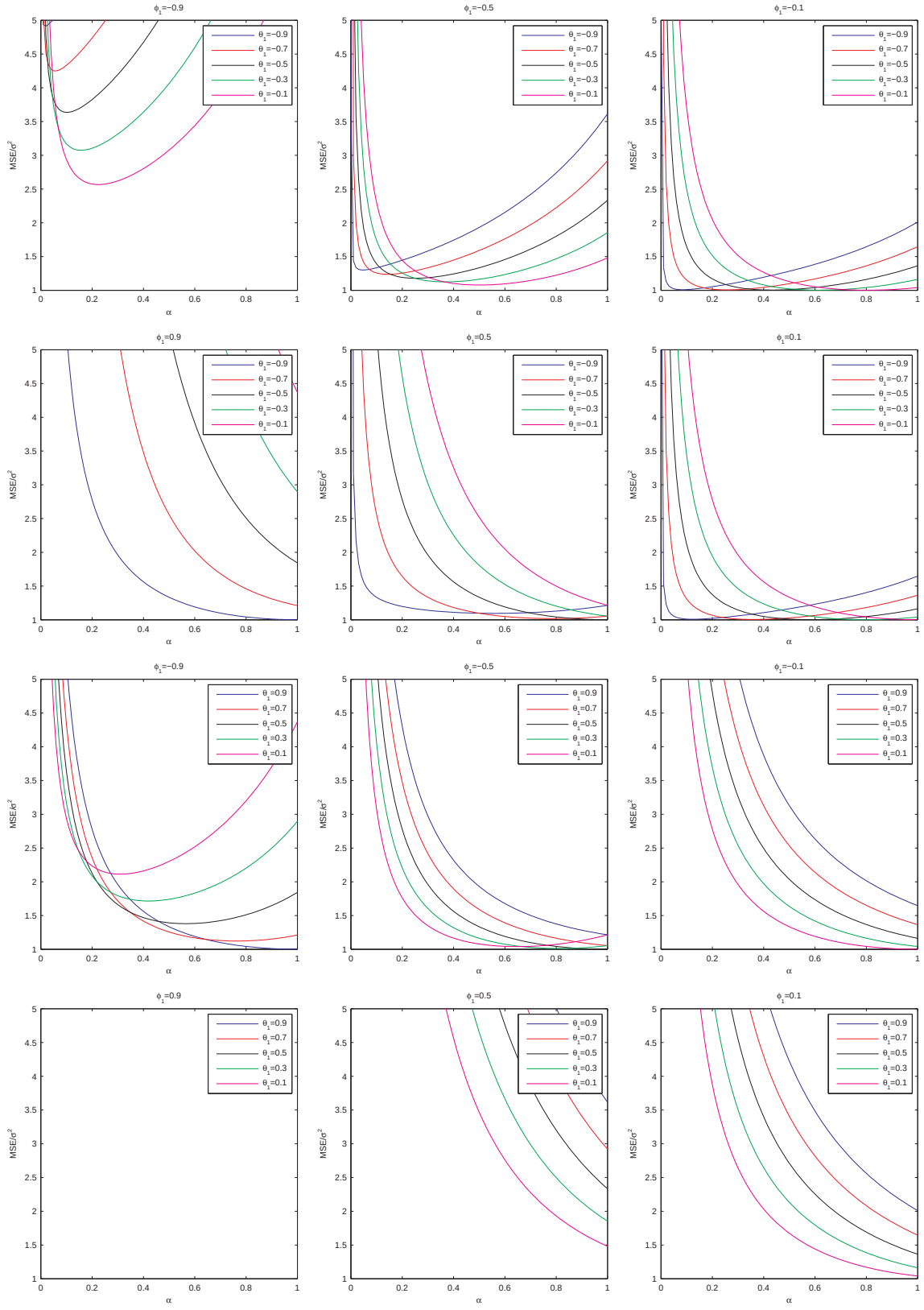


Figure 3.9: **SES** – MSE/σ^2 as a function of α, N_t is an ARIMA(1,1,1)

Table 3.8: SES – Optimal α and Minimum MSE/σ^2, N_t is an ARIMA(1,1,1)

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
-0.90	0.02	4.91	0.05	1.30	0.08	1.01
-0.70	0.06	4.25	0.14	1.24	0.26	1.01
-0.50	0.10	3.63	0.25	1.18	0.43	1.00
-0.30	0.16	3.07	0.37	1.12	0.61	1.00
-0.10	0.22	2.57	0.50	1.08	0.80	1.00

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
-0.90	1.00	1.00	0.56	1.09	0.12	1.01
-0.70	1.00	1.21	0.79	1.02	0.36	1.00
-0.50	1.00	1.84	1.00	1.00	0.58	1.00
-0.30	1.00	2.89	1.00	1.05	0.80	1.00
-0.10	1.00	4.37	1.00	1.21	1.00	1.00

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
0.90	1.00	1.00	1.00	1.21	1.00	1.65
0.70	0.76	1.12	1.00	1.05	1.00	1.36
0.50	0.57	1.38	1.00	1.00	1.00	1.16
0.30	0.42	1.71	0.81	1.01	1.00	1.04
0.10	0.31	2.11	0.65	1.04	1.00	1.00

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
0.90	1.00	18.05	1.00	3.61	1.00	2.01
0.70	1.00	14.47	1.00	2.92	1.00	1.65
0.50	1.00	11.32	1.00	2.33	1.00	1.36
0.30	1.00	8.58	1.00	1.85	1.00	1.16
0.10	1.00	6.26	1.00	1.48	1.00	1.04

Therefore, $\psi_0 = 1$, $\psi_1 = \phi_1 + \theta_1 - 1$, and $\psi_i = (\phi_1 + \theta_1)(\phi_1 - 1)\phi_1^{i-2}$ for $i \geq 2$. The MSE becomes

$$\begin{aligned} E[e_t^2] &= \sigma^2[1 + \sum_{j=1}^{\infty}(\frac{\lambda^{j-1}(\lambda+\theta_1)(\lambda-1) - \phi_1^{j-1}(\phi_1+\theta_1)(\phi_1-1)}{\lambda - \phi_1})^2] \\ &= \sigma^2 \cdot [\frac{2\theta_1}{1-\lambda\phi_1} + \frac{2(\theta_1-1)^2}{(1+\lambda)(1+\phi_1)(1-\lambda\phi_1)}]. \end{aligned} \tag{3.41}$$

Figure 3.10 and Table 3.9 show that

- SES performs well when $\theta_1\phi_1 < 0$ and $|\theta_1|$ and $|\phi_1|$ are close.
- α_{opt} increases as both θ_1 and ϕ_1 increase. $\alpha_{opt} \to 0$ when either $-1 < \theta_1 < 0$ and $|\phi_1|$ is small or $-1 < \phi_1 < 0$.
- When $0 < \theta_1 < 1$, the smaller the $|\phi_1|$, the less critical the choice of α.

3.2.5 Summary

Based on the results above on the performance of SES for different types of ARIMA time series, the following conclusions can be draw.

- N_t is an ARIMA$(0, 1, q)$, $0 \leq q \leq 2$. $\alpha_{opt} = 1$ when $\theta_1 > 0$, $\alpha_{opt} < 1$ when $\theta_1 < 0$, and overestimation of α is less serious than the equivalent underestimation. SES performs well when $\theta_1 < 0$ and $|\theta_2|$ is small.
- N_t is an ARIMA$(0, 0, q)$, $0 \leq q \leq 2$. α_{opt} is often extremely small ($\to 0$), and the choice of α is not critical when $\theta_1 > 0$ and $|\theta_2|$ is small. SES performs well when both $|\theta_1|$ and $|\theta_2|$ are small.
- N_t is an ARIMA$(1, 1, 0)$. $\alpha_{opt} = 1$ when $0 < \phi_1 < 1$, $\alpha_{opt} < 1$ when $-1 < \phi_1 < 0$, and overestimation of α is less serious than the equivalent underestimation. SES performs well when $|\phi_1|$ is small, say $-0.7 < \phi_1 < 0.5$.

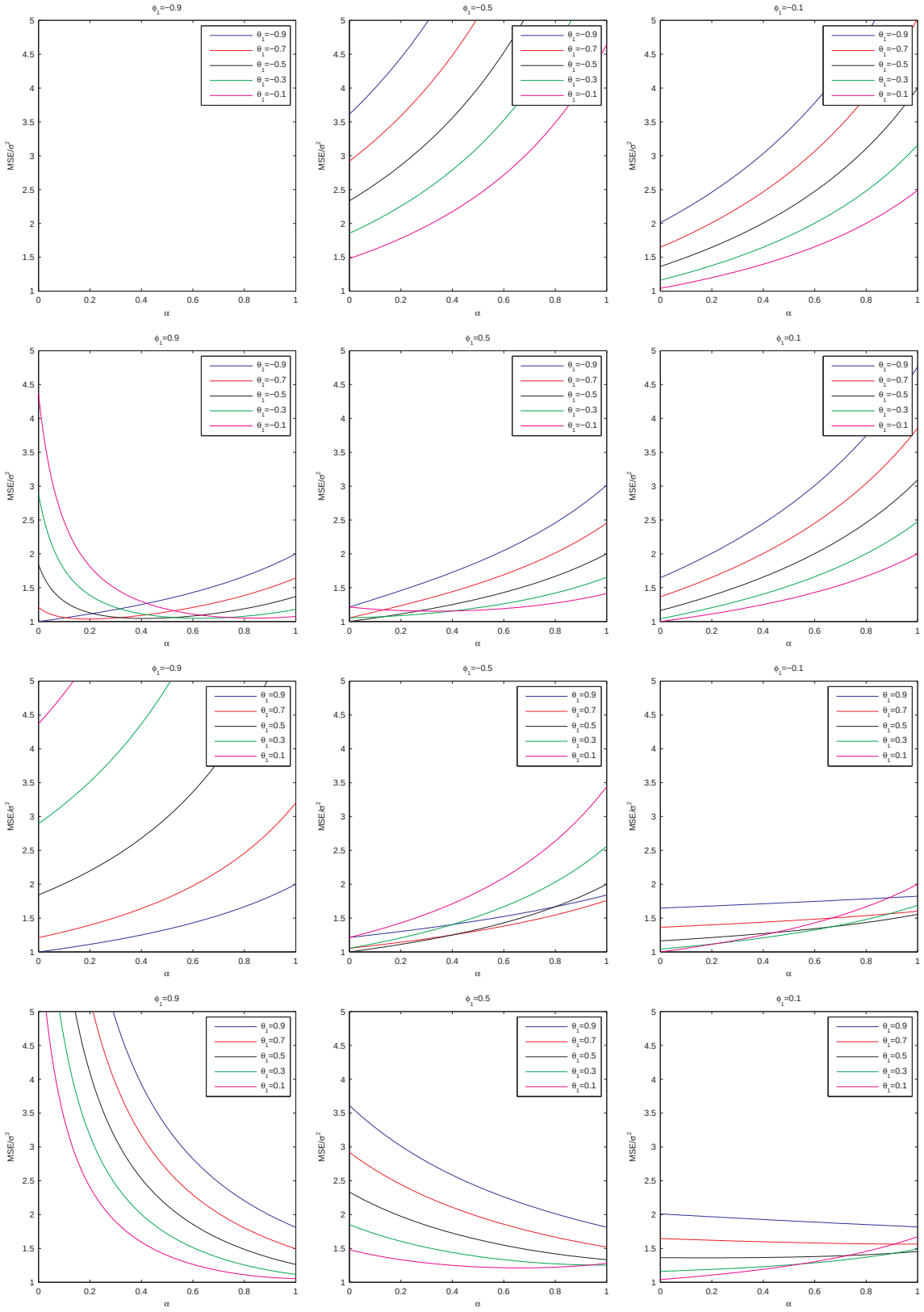


Figure 3.10: SES – MSE/σ^2 as a function of α, N_t is an ARMA(1,1)

Table 3.9: SES – Optimal α and Minimum MSE/σ^2, N_t is an ARMA(1,1)

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
-0.90	$\to 0$	18.05	$\to 0$	3.61	$\to 0$	2.01
-0.70	$\to 0$	14.47	$\to 0$	2.92	$\to 0$	1.65
-0.50	$\to 0$	11.32	$\to 0$	2.33	$\to 0$	1.36
-0.30	$\to 0$	8.58	$\to 0$	1.85	$\to 0$	1.16
-0.10	$\to 0$	6.26	$\to 0$	1.48	$\to 0$	1.04

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
-0.90	$\to 0$	1.00	$\to 0$	1.21	$\to 0$	1.65
-0.70	0.19	1.04	$\to 0$	1.05	$\to 0$	1.36
-0.50	0.41	1.04	$\to 0$	1.00	$\to 0$	1.16
-0.30	0.63	1.05	$\to 0$	1.05	$\to 0$	1.04
-0.10	0.84	1.05	0.33	1.15	$\to 0$	1.00

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
0.90	$\to 0$	1.00	$\to 0$	1.21	$\to 0$	1.65
0.70	$\to 0$	1.21	$\to 0$	1.05	$\to 0$	1.36
0.50	$\to 0$	1.84	$\to 0$	1.00	$\to 0$	1.16
0.30	$\to 0$	2.89	$\to 0$	1.05	$\to 0$	1.04
0.10	$\to 0$	4.37	$\to 0$	1.21	$\to 0$	1.00

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2	α_{opt}	MSE/σ^2
0.90	1.00	1.81	1.00	1.81	1.00	1.82
0.70	1.00	1.49	1.00	1.52	0.98	1.56
0.50	1.00	1.26	1.00	1.33	0.17	1.36
0.30	1.00	1.12	0.98	1.25	$\to 0$	1.16
0.10	1.00	1.05	0.67	1.21	$\to 0$	1.04

- N_t is an ARIMA$(1,0,0)$. $\alpha_{opt} \to 0$ when $-1 < \phi_1 \leq 1/3$, $\alpha_{opt} = (3\phi_1 - 1)/2\phi_1$ when $1/3 < \phi_1 < 1$, and the choice of α is not critical when ϕ_1 is close to 0.5. SES performs well ϕ_1 is large, say $-0.5 < \phi_1 < 1$.

- N_t is an ARIMA$(1,1,1)$. α_{opt} increases as θ_1 and/or ϕ_1 increase, and overestimation of α is less serious than the equivalent underestimation. SES performs well when either $-1 < \theta_1 < 0$ and $|\phi_1|$ is small or $\theta_1\phi_1 < 0$ and $|\theta_1| \approx |\phi_1|$

- N_t is an ARIMA$(1,0,1)$. α_{opt} is often extremely small ($\to 0$) and increases as θ_1 and/or ϕ_1 increase. When $0 < \theta_1 < 1$, the smaller the $|\phi_1|$, the less critical the choice of α. SES performs well when $\theta_1\phi_1 < 0$ and $|\theta_1| \approx |\phi_1|$.

As a result, when N_t is an ARIMA$(p,1,q)$ with $0 \leq p \leq 1$ and $0 \leq q \leq 2$, α_{opt} tends to be large, and overestimation of α is less serious than the equivalent underestimation. In addition, SES performs well for an IMA$(1,q)$ with $\theta_1 < 0$ and $|\theta_2|$ small, an ARI$(1,1)$ with small $|\phi_1|$, and an ARIMA$(1,1,1)$ with $-1 < \theta_1 < 0$ and $|\phi_1|$ small . When N_t is an ARMA(p,q) with $0 \leq p \leq 1$ and $0 \leq q \leq 2$, α_{opt} tends to be small, and the choice of α is often not critical. In addition, SES performs well for an MA(q) with small $|\theta_1|$ and $|\theta_2|$, an AR(1) with large ϕ_1, and an ARMA$(1,1)$ with $\theta_1\phi_1 < 0$ and $|\theta_1| \approx |\phi_1|$.

3.3 Holt's Method

The one-step-ahead forecast by Holt's method is

$$\hat{Y}_{t|t-1} = l_{t-1} + b_{t-1}, \tag{3.42}$$

where l_t and b_t are updated as follow

$$l_t = \alpha_1 Y_t + (1-\alpha_1)(l_{t-1} + b_{t-1}), \tag{3.43a}$$

$$b_t = \alpha_2(l_t - l_{t-1}) + (1-\alpha_2)b_{t-1}, \tag{3.43b}$$

where α_1 and α_2 are smoothing parameters taking values in the interval (0,1]. As in SES, $\hat{Y}_{t|t-1}$ can be expressed in terms of past observations $Y_{t-k}, k > 0$, and past one-step-ahead forecasts $\hat{Y}_{t-k}(1), k > 0$,

$$\hat{Y}_{t|t-1} = (\alpha_1 + \alpha_1\alpha_2)Y_{t-1} - \alpha_1 Y_{t-2} + (2 - \alpha_1 - \alpha_1\alpha_2)\hat{Y}_{t-1|t-2} - (1-\alpha_1)\hat{Y}_{t-2|t-3} \tag{3.44}$$

Assume that the true data generating process is

$$Y_t = \mu_0 + \delta t + N_t, \tag{3.45}$$

where μ_0 and δ are constant, and N_t is a disturbance term with zero mean and second difference as a stationary process

$$(1-B)^2 N_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}, \tag{3.46}$$

Holt's method with smoothing parameters α_1 and α_2 for model (3.45) leads to

$$e_t - (2 - \alpha_1 - \alpha_1\alpha_2)e_{t-1} - (\alpha_1 - 1)e_{t-2} = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}. \tag{3.47}$$

Let λ_1 and λ_2 be the roots of the equation

$$x^2 - (2 - \alpha_1 - \alpha_1\alpha_2)x - (\alpha_1 - 1) = 0. \tag{3.48}$$

The fact that $0 < \alpha_1 \leq 1$ and $0 < \alpha_2 \leq 1$ implies that

$$|\alpha_1 - 1| < 1, \tag{3.49a}$$

$$(\alpha_1 - 1) + (2 - \alpha_1 - \alpha_1\alpha_2) = 1 - \alpha_1\alpha_2 < 1, \tag{3.49b}$$

$$(\alpha_1 - 1) - (2 - \alpha_1 - \alpha_1\alpha_2) = 2\alpha_1 + \alpha_1\alpha_2 - 3 < 1, \tag{3.49c}$$

which are necessary and sufficient conditions for the roots λ_1 and λ_2 to be less than 1 in absolute value. As a result, e_t can be written as

$$\begin{aligned} e_t &= \frac{1}{(1-\lambda_1 B)(1-\lambda_2 B)} \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} \\ &= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \lambda_1^j \lambda_2^k \psi_i \epsilon_{t-i-j-k} \\ &= \sum_{j=0}^{\infty} (\sum_{i=0}^{j} \psi_i \sum_{k=0}^{j-i} \lambda_1^{j-i-k} \lambda_2^k) \epsilon_{t-j}. \end{aligned} \tag{3.50}$$

As a result, the mean of the one-step-ahead forecast error is

$$E[e_t] = \sum_{j=0}^{\infty} (\sum_{i=0}^{j} \psi_i \sum_{k=0}^{j-i} \lambda_1^{j-i-k} \lambda_2^k) E[\epsilon_{t-j}] = 0, \tag{3.51}$$

which implies that the one-step-ahead forecast by Holt's method for model (3.45) is unbiased. The mean squared one-step-ahead forecast error is

$$E[e_t^2] = \sigma^2 \sum_{j=0}^{\infty} (\sum_{i=0}^{j} \psi_i \sum_{k=0}^{j-i} \lambda_1^{j-i-k} \lambda_2^k)^2. \tag{3.52}$$

3.3.1 N_t is an ARIMA$(0, 2, q)$ process

N_t is an ARIMA$(0, 2, q)$ process

$$(1-B)^2 N_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i}. \tag{3.53}$$

Therefore, $\psi_i = \theta_i$ for $0 \le i \le q$, and $\psi_i = 0$ for $i > q$. The MSE becomes

$$E[e_t^2] = \sigma^2 [\sum_{j=0}^{q-1} (\sum_{i=0}^{j} \theta_i \sum_{k=0}^{j-i} \lambda_1^{j-i-k} \lambda_2^k)^2 + \sum_{j=q}^{\infty} (\sum_{i=0}^{q} \theta_i \sum_{k=0}^{j-i} \lambda_1^{j-i-k} \lambda_2^k)^2]. \tag{3.54}$$

1). $q = 0$, N_t is an ARIMA(0,2,0).

$$E[e_t^2] = \sigma^2 \cdot \frac{1+\lambda_1\lambda_2}{(1-\lambda_1^2)(1-\lambda_2^2)(1-\lambda_1\lambda_2)}. \tag{3.55}$$

Figure 3.11 shows that the MSE is a monotone decreasing function of both α_1 and α_2. The minimum MSE occurs as $\alpha_1 = \alpha_2 = 1$, at which the one-step-ahead forecast given by Holt's method is

$$\hat{Y}_{t|t-1} = 2Y_{t-1} - Y_{t-2}. \tag{3.56}$$

Such a result is due to the ARIMA(0,2,0) disturbance, which has the property

$$E[Y_t|Y_{t-1}, Y_{t-2}, \cdots] = E[Y_t|Y_{t-1}, Y_{t-2}] = 2Y_{t-1} - Y_{t-2}. \tag{3.57}$$

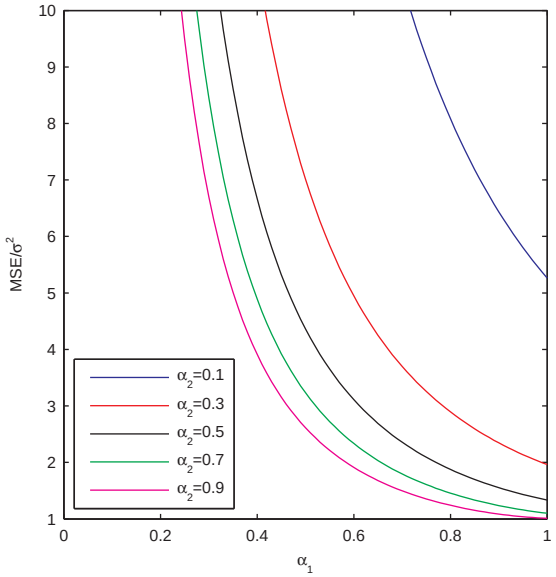


Figure 3.11: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(0,2,0)

2). $q = 1$, N_t is an ARIMA(0,2,1).

$$E[e_t^2] = \sigma^2 \cdot \frac{(\lambda_1 + \theta_1)(\lambda_2 + \theta_1) + (1 + \lambda_1\theta_1)(1 + \lambda_2\theta_1)}{(1 - \lambda_1^2)(1 - \lambda_2^2)(1 - \lambda_1\lambda_2)}. \tag{3.58}$$

Figure 3.12 and Table 3.10 shows that

- Holt's method performs better when $-1 < \theta_1 < 0$ than when $0 < \theta_1 < 1$. When $-1 < \theta_1 < 0$, the minimum MSE stays at σ^2, the theoretically optimal value, regardless of what value θ_1 takes. When $0 < \theta_1 < 1$, the minimum MSE increases as θ_1 increases.
- The optimal value of α_1 stays at 1, and the optimal value of α_2 equals to $1 + \theta_1$ when $-1 < \theta_1 < 0$ and stays at 1 when $0 < \theta_1 < 1$. When

$\alpha_1 = \alpha_2 = 1$, according to equation (3.58), the minimum MSE equals to $\sigma^2(1+\theta_1^2)$.

In fact, when $-1 < \theta_1 < 0$, values of α_1 and α_2 can be found so that Holt's method gives optimal forecasts. For an ARIMA(0,2,1) disturbance, equation (3.47) becomes

$$e_t - (2 - \alpha_1 - \alpha_1\alpha_2)e_{t-1} - (\alpha_1 - 1)e_{t-2} = \epsilon_t + \theta_1\epsilon_{t-1}, \tag{3.59}$$

which suggests that, when $\alpha_1 = 1$ and $\alpha_2 = 1 + \theta_1$, e_t becomes a white noise. Therefore, the forecast by Holt's method is optimal.

– The smaller the θ_1, the less critical the choice of α_2, and overestimation of α_2 is less serious than the equivalent underestimation.

In summary, Holt's method performs well when θ_1 is small. Large α_1 and α_2 are preferred.

Table 3.10: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARIMA(0,2,1)

θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.95	(1.00, 0.05)	1.00	0.95	(1.00, 1.00)	1.90
-0.90	(1.00, 0.10)	1.00	0.90	(1.00, 1.00)	1.81
-0.80	(1.00, 0.20)	1.00	0.80	(1.00, 1.00)	1.64
-0.70	(1.00, 0.30)	1.00	0.70	(1.00, 1.00)	1.49
-0.60	(1.00, 0.40)	1.00	0.60	(1.00, 1.00)	1.36
-0.50	(1.00, 0.50)	1.00	0.50	(1.00, 1.00)	1.25
-0.40	(1.00, 0.60)	1.00	0.40	(1.00, 1.00)	1.16
-0.30	(1.00, 0.70)	1.00	0.30	(1.00, 1.00)	1.09
-0.20	(1.00, 0.80)	1.00	0.20	(1.00, 1.00)	1.04
-0.10	(1.00, 0.90)	1.00	0.10	(1.00, 1.00)	1.01

3). $q = 2$, N_t is an ARIMA(0,2,2).

Figures 3.13 and 3.14 and Table 3.11 show that

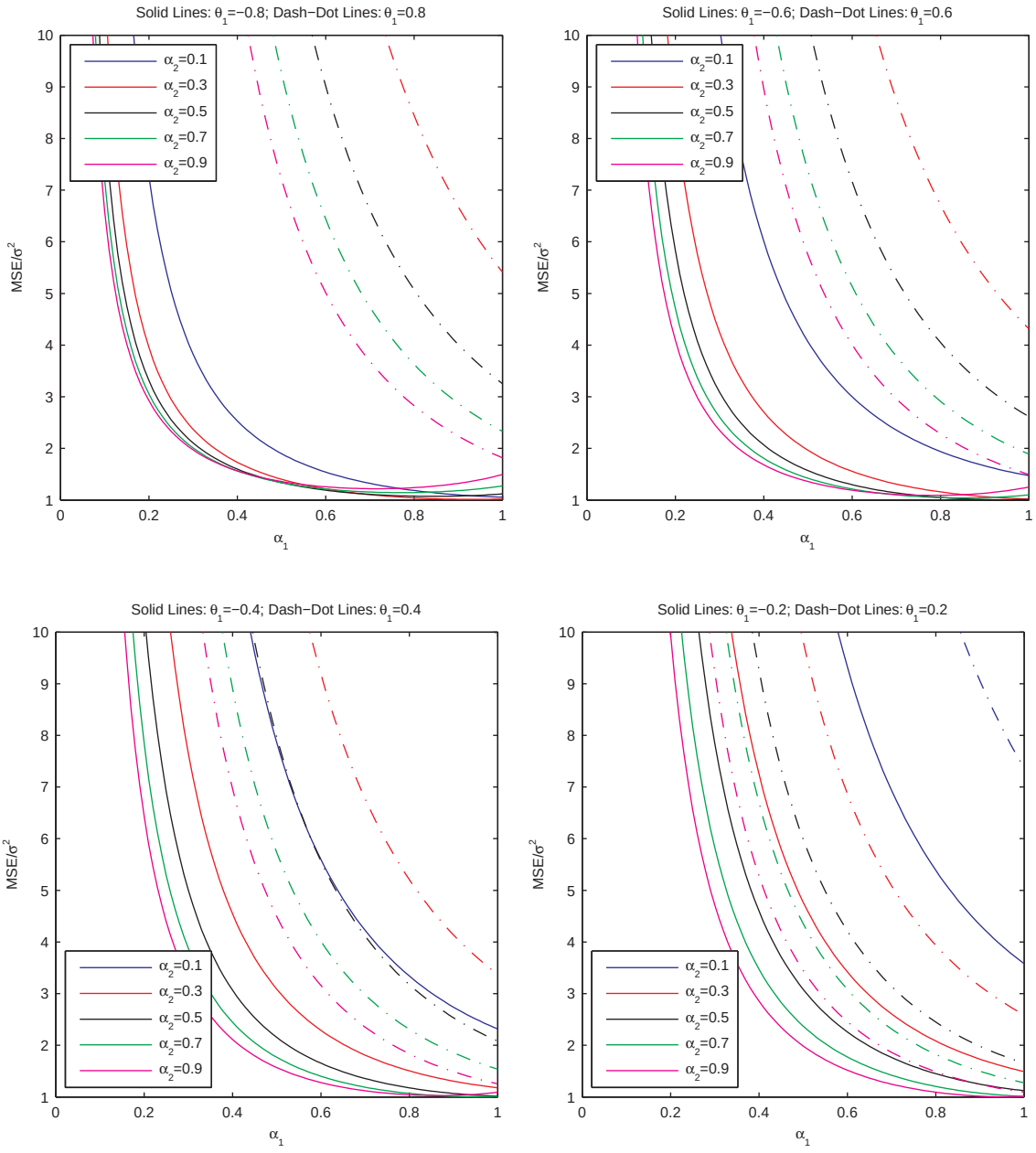


Figure 3.12: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(0,2,1)

- When $0 < \theta_2 < 1$, the minimum MSE stays at σ^2 for small θ_1 and then increases as θ_1 and/or θ_2 increase.
- When $-1 < \theta_2 < 0$ and $\theta_1 \leq 0$, the value of θ_1 has no effect on the minimum MSE and the minimum MSE increases as θ_2 decreases. When $-1 < \theta_2 < 0$ and $\theta_1 > 0$, the minimum MSE increases as θ_1 increases while decreases as θ_2 increases.
- When $0 < \theta_2 < 1$, the optimal value of α_1, starting from $1 - \theta_2$, increases with θ_1 once $\theta_1 > -2\theta_2$ before reaching 1. In addition, overestimation of α_1 is less serious than the equivalent underestimation.
- When $-1 < \theta_2 < 0$, the optimal value of α_1 stays at 1, the optimal value of α_2 increases with θ_1 before reaching 1, and the smaller the θ_1, the less critical the choice of α_2.

In fact, Holt's method is optimal for an ARIMA(0,2,2) when

$$\alpha_1 = 1 - \theta_2 \text{ and } \alpha_2 = \frac{1 + \theta_1 + \theta_2}{1 - \theta_2} \tag{3.60}$$

The fact that $\alpha_1 \in (0, 1]$ and $\alpha_2 \in (0, 1]$ implies that only for an ARIMA(0,2,2) with θ_1 and θ_2 falling inside the region formed by (see Figure 2.1 in Chapter 2)

$$-(1 + \theta_2) < \theta_1 \leq -2\theta_2 \text{ and } \quad 0 \leq \theta_2 < 1, \tag{3.61}$$

could Holt's method be optimal.

In summary, Holt's method is optimal for an ARIMA(0,2,2) when θ_1 and θ_2 satisfy equation (3.60). Large α_1 and α_2 are perferred.

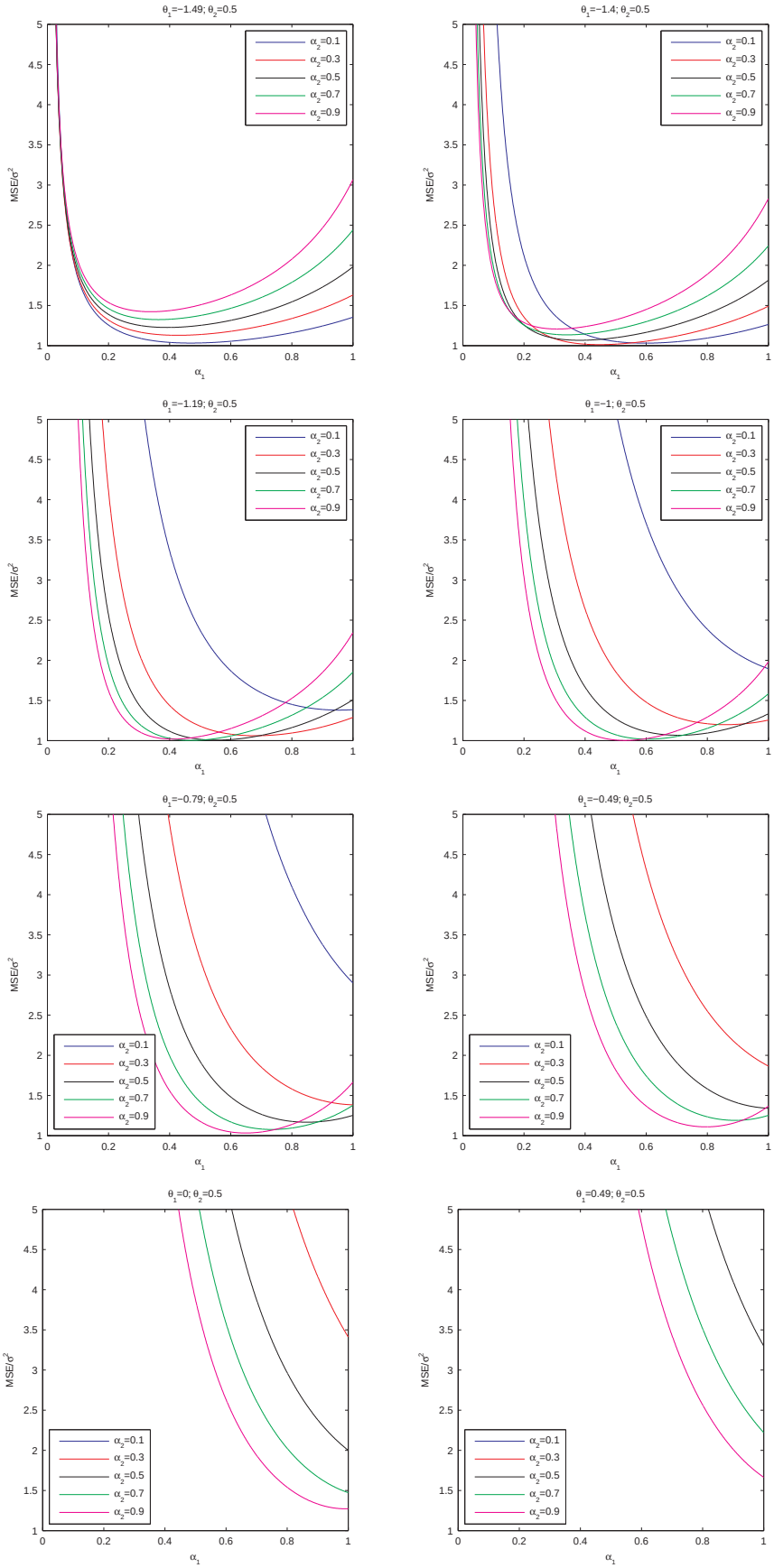


Figure 3.13: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(0,2,2)

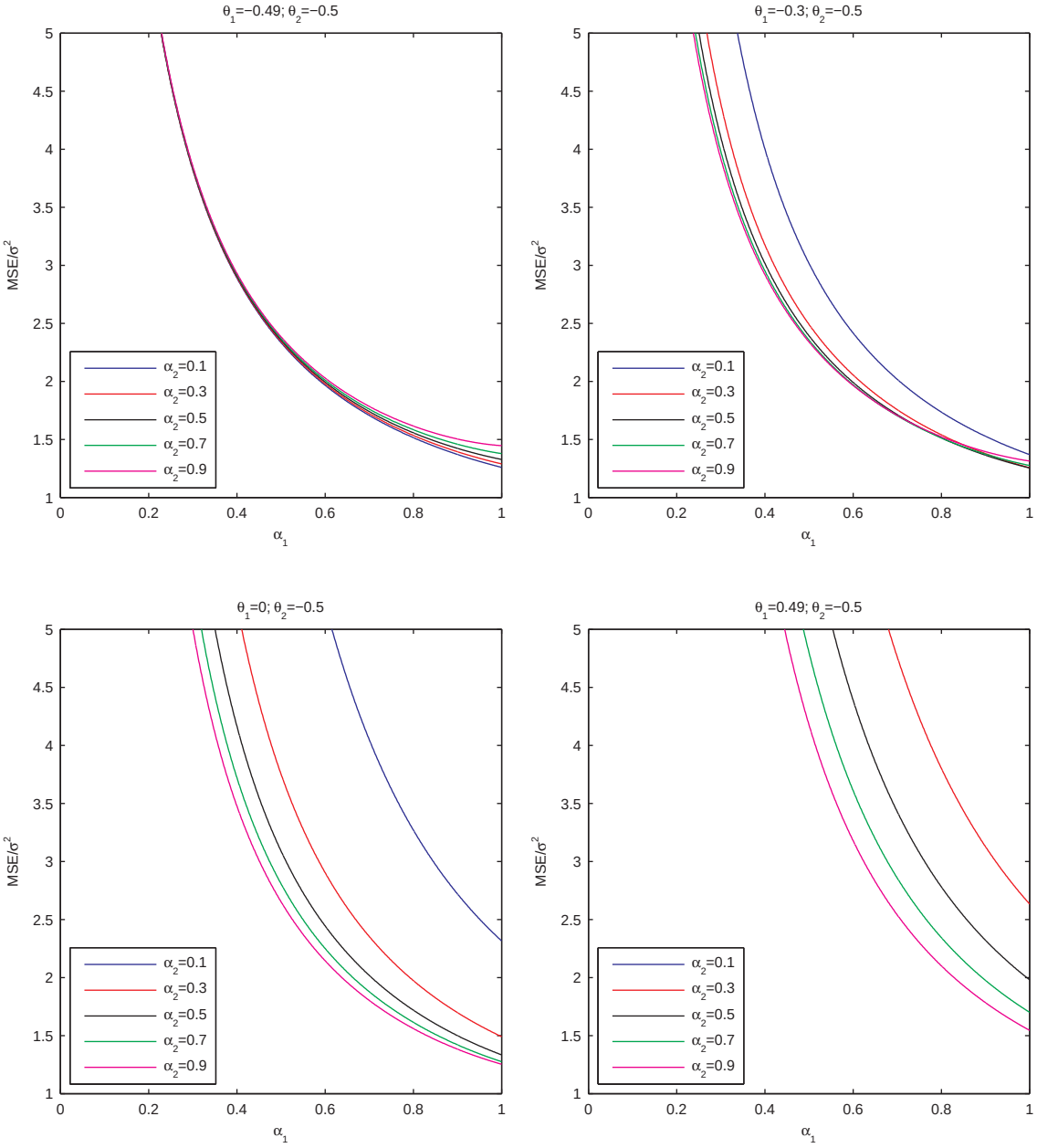


Figure 3.14: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(0,2,2)

Table 3.11: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARIMA(0,2,2)

	$\theta_2 = -0.2$		$\theta_2 = 0.2$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-1.19	-	-	(0.80, 0.01)	1.00
-1.00	-	-	(0.80, 0.25)	1.00
-0.79	(1.00, 0.01)	1.04	(0.80, 0.51)	1.00
-0.60	(1.00, 0.25)	1.04	(0.80, 0.75)	1.00
-0.49	(1.00, 0.39)	1.04	(0.80, 0.89)	1.00
-0.40	(1.00, 0.50)	1.04	(0.80, 1.00)	1.00
-0.20	(1.00, 0.75)	1.04	(0.89, 1.00)	1.01
0.00	(1.00, 1.00)	1.04	(0.97, 1.00)	1.03
0.20	(1.00, 1.00)	1.08	(1.00, 1.00)	1.08
0.40	(1.00, 1.00)	1.20	(1.00, 1.00)	1.20
0.49	(1.00, 1.00)	1.28	(1.00, 1.00)	1.28
0.60	(1.00, 1.00)	1.40	(1.00, 1.00)	1.40
0.79	(1.00, 1.00)	1.66	(1.00, 1.00)	1.66
1.00	-	-	(1.00, 1.00)	2.04
1.19	-	-	(1.00, 1.00)	2.46
	$\theta_2 = -0.5$		$\theta_2 = 0.5$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-1.49	-	-	(0.50, 0.02)	1.00
-1.40	-	-	(0.50, 0.20)	1.00
-1.19	-	-	(0.50, 0.62)	1.00
-1.00	-	-	(0.50, 1.00)	1.00
-0.79	-	-	(0.61, 1.00)	1.02
-0.60	-	-	(0.71, 1.00)	1.06
-0.49	(1.00, 0.02)	1.25	(0.76, 1.00)	1.08
-0.40	(1.00, 0.20)	1.25	(0.79, 1.00)	1.11
-0.20	(1.00, 0.60)	1.25	(0.87, 1.00)	1.16
0.00	(1.00, 1.00)	1.25	(0.94, 1.00)	1.22
0.20	(1.00, 1.00)	1.29	(1.00, 1.00)	1.29
0.40	(1.00, 1.00)	1.41	(1.00, 1.00)	1.41
0.49	(1.00, 1.00)	1.49	(1.00, 1.00)	1.49
0.60	-	-	(1.00, 1.00)	1.61
0.79	-	-	(1.00, 1.00)	1.87
1.00	-	-	(1.00, 1.00)	2.25
1.19	-	-	(1.00, 1.00)	2.67

3.3.2 N_t is an ARIMA$(0,1,q)$ process

N_t is an ARIMA$(0,1,q)$ process

$$(1-B)N_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i}. \tag{3.62}$$

Let $\theta_{-1} = \theta_{q+1} = 0$, then

$$(1-B)^2 N_t = \sum_{i=0}^{q+1} (\theta_i - \theta_{i-1}) \epsilon_{t-i}. \tag{3.63}$$

1). $q = 0$, N_t is an ARIMA(0,1,0), which is a random walk.

$$E[e_t^2] = \sigma^2 \cdot \frac{2}{(1+\lambda_1)(1+\lambda_2)(1-\lambda_1\lambda_2)}. \tag{3.64}$$

Figure 3.15 shows that a large α_1 and a small α_2 give a small MSE. The minimum MSE occurs as $\alpha_1 = 1$ and $\alpha_2 \to 0$. Also, the choice of α_2 is not as critical as the choice of α_1.

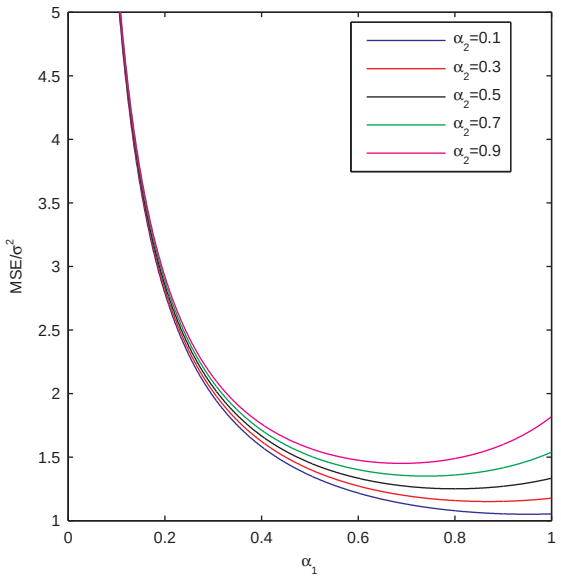


Figure 3.15: Holt's Method – MSE/σ^2 as a function of α_1, N_t is a random walk

2). $q = 1$, N_t is an ARIMA(0,1,1).

Figure 3.16 and Table 3.12 show that

– Holt's method performs better when $-1 < \theta_1 < 0$ than when $0 < \theta_1 < 1$. When $-1 < \theta_1 < 0$, the minimum MSE stays at σ^2 no matter what value θ_1 takes. When $0 < \theta_1 < 1$, the minimum MSE increases as θ_1 increases.

– When $-1 < \theta_1 < 0$, the minimum MSE occurs as $\alpha_1 = 1 + \theta_1$ and $\alpha_2 \to 0$, and overestimation of α_1 is less serious than the equivalent underestimation. When $0 < \theta_1 < 1$, the minimum MSE occurs as $\alpha_1 = 1$ and $\alpha_2 \to 0$.

– The larger the θ_1, the less critical the choice of α_2.

In summary, Holt's method performs well for an ARIMA(0,1,1) when $-1 < \theta_1 < 0$, and a small α_2 is preferred.

Table 3.12: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARIMA(0,1,1)

θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.95	(0.05, $\to$ 0)	1.00	0.95	(1.00, $\to$ 0)	1.90
-0.90	(0.10, $\to$ 0)	1.00	0.90	(1.00, $\to$ 0)	1.81
-0.80	(0.20, $\to$ 0)	1.00	0.80	(1.00, $\to$ 0)	1.64
-0.70	(0.30, $\to$ 0)	1.00	0.70	(1.00, $\to$ 0)	1.49
-0.60	(0.40, $\to$ 0)	1.00	0.60	(1.00, $\to$ 0)	1.36
-0.50	(0.50, $\to$ 0)	1.00	0.50	(1.00, $\to$ 0)	1.25
-0.40	(0.60, $\to$ 0)	1.00	0.40	(1.00, $\to$ 0)	1.16
-0.30	(0.70, $\to$ 0)	1.00	0.30	(1.00, $\to$ 0)	1.09
-0.20	(0.80, $\to$ 0)	1.00	0.20	(1.00, $\to$ 0)	1.04
-0.10	(0.90, $\to$ 0)	1.00	0.10	(1.00, $\to$ 0)	1.01

3). $q = 2$, N_t is an ARIMA(0,1,2).

Figures 3.17 and 3.18 and Table 3.13 show that

– When $\theta_1 \leq 0$, the minimum MSE depends only on $|\theta_2|$, the absolute value of θ_2, and is independent of both the sign of θ_2 and the value of θ_1. Moreover, the larger the $|\theta_2|$, the larger the minimum MSE.

– When $\theta_1 > 0$, the minimum MSE increases as θ_1 increases.

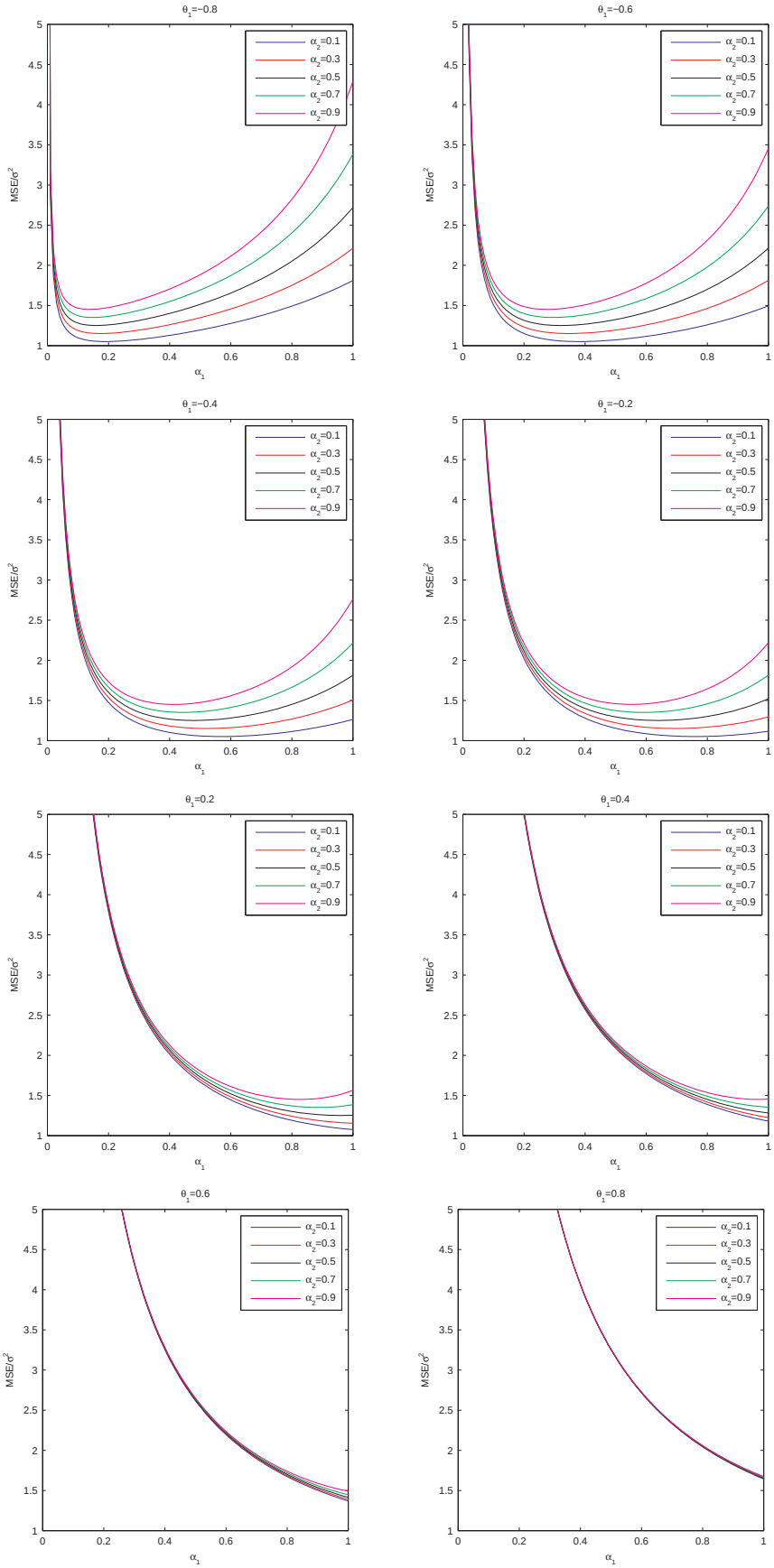


Figure 3.16: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(0,1,1)

- When $\theta_1 \leq 0$, the optimal value of α_1 increases as θ_1 increases and reaches 1 at $\theta_1 = 0$, while the optimal value of α_2 stays at an extremely small value ($\rightarrow 0$).
- When $\theta_1 > 0$, the optimal value of α_1 stays at 1, while the optimal value of α_2 is extremely small ($\rightarrow 0$) when $-1 < \theta_2 < 0$ and increases with θ_1 when $0 < \theta_2 < 1$.
- The smaller the θ_2, the less critical the choice of α_1.

In summary, Holt's method performs well for an ARIMA(0,1,2) when $\theta_1 < 0$ and $|\theta_2|$ is small, and a small α_2 is preferred.

Comparing the results for ARIMA(0,1,q) from SES (Tables 3.1 and 3.2) and Holt's method (Tables 3.12 and 3.13) respectively reveals that Holt's method with $\alpha_2 \rightarrow 0$ gives exactly the same results as these by SES.

When $\alpha_2 \rightarrow 0$, the updating equation for b_t in Holt's method becomes $b_t \approx b_{t-1}$. That is, there is no updating for the slope estimate b_t anymore. Two situations where no updating for b_t is necessary are i) the slope δ in Equation (3.45) is known; ii) a fairly good estimate of δ is already available and including new data improves very little. Under such situations, Holt's method reduces to

$$l_t = \alpha_1 Y_t + (1 - \alpha_1)(l_{t-1} + \delta) \tag{3.65}$$

and the one-step-ahead forecast by Holt's method becomes

$$\hat{Y}_{t|t-1} = \alpha_1 Y_{t-1} + (1 - \alpha_1)\hat{Y}_{t-1|t-2} + \delta \tag{3.66}$$

which looks like the one-step-ahead forecast by SES with an extra term δ. Also, the one-step-ahead forecast errors from Holt's method now satisfy

$$e_t - (1 - \alpha_1)e_{t-1} = (1 - B)N_t \tag{3.67}$$

Therefore, as long as $(1 - B)N_t$ is stationary, Holt's method with $\alpha_2 \rightarrow 0$ for (3.45) will give exactly the same results as these by SES for (3.2).

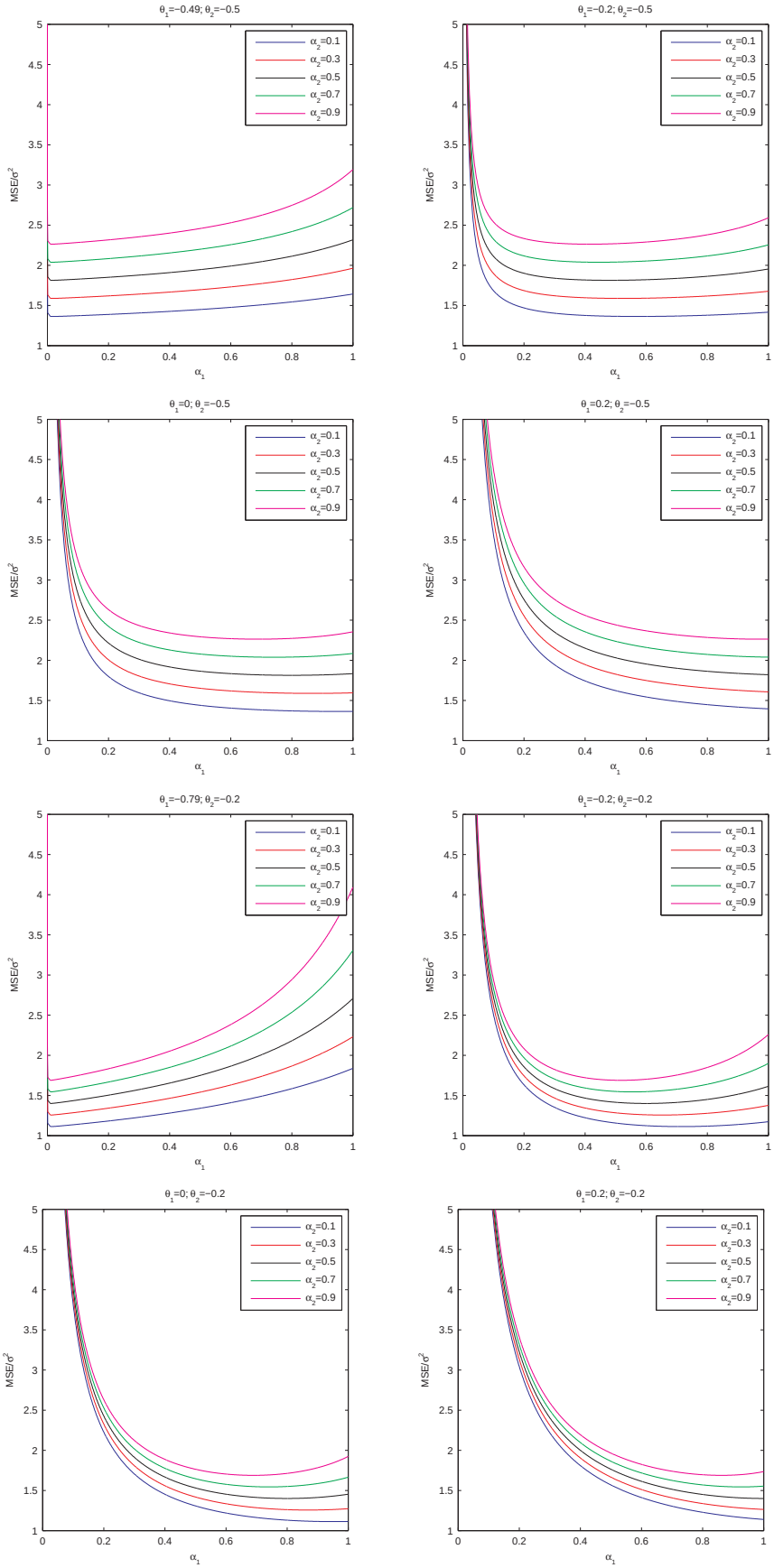


Figure 3.17: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(0,1,2)

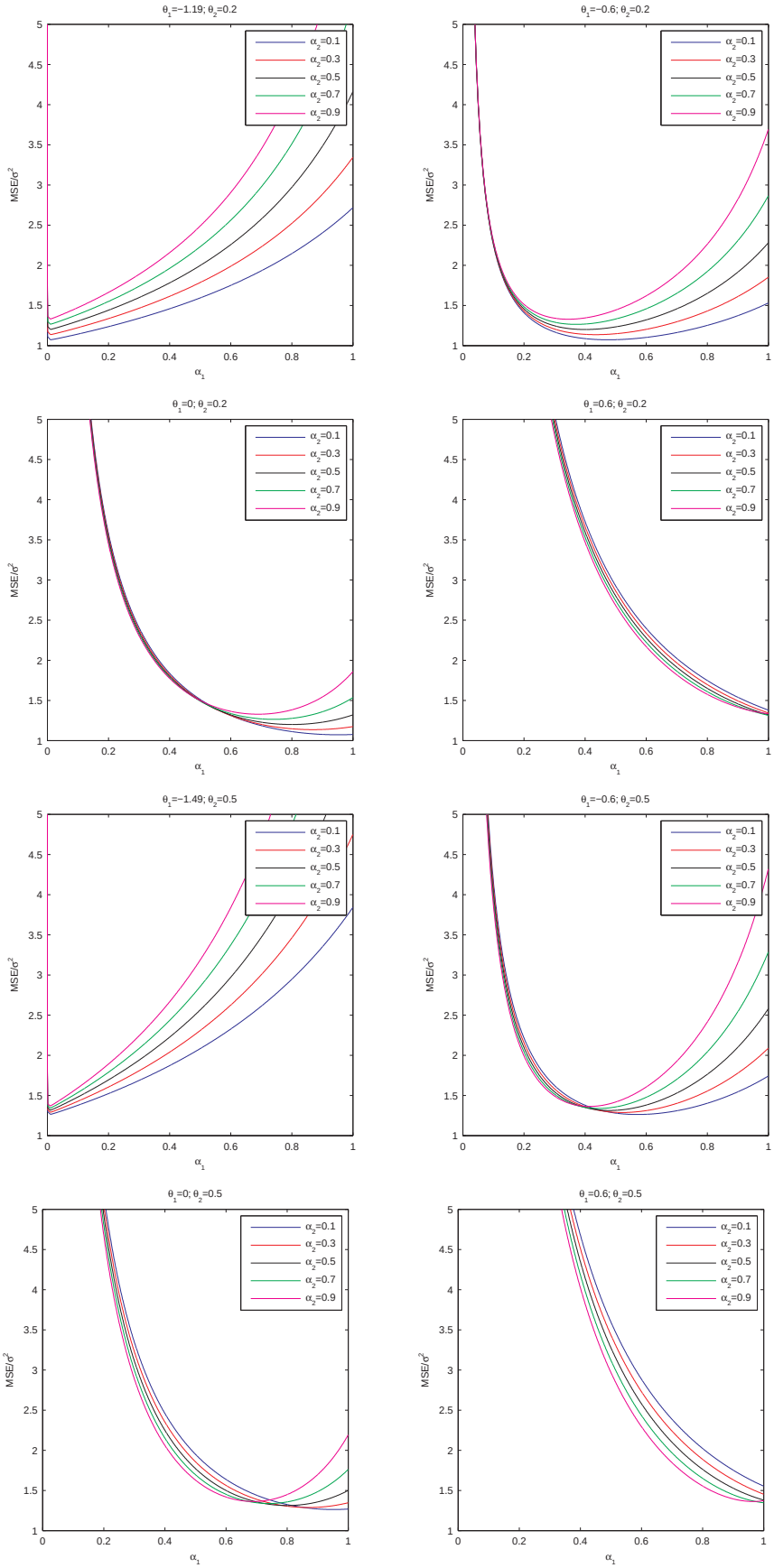


Figure 3.18: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(0,1,2)

Table 3.13: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARIMA(0,1,2)

	$\theta_2 = -0.2$		$\theta_2 = 0.2$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-1.19	-	-	(0.01, $\rightarrow$ 0)	1.04
-1.00	-	-	(0.17, $\rightarrow$ 0)	1.04
-0.79	(0.01, $\rightarrow$ 0)	1.04	(0.34, $\rightarrow$ 0)	1.04
-0.60	(0.25, $\rightarrow$ 0)	1.04	(0.50, $\rightarrow$ 0)	1.04
-0.49	(0.39, $\rightarrow$ 0)	1.04	(0.59, $\rightarrow$ 0)	1.04
-0.40	(0.50, $\rightarrow$ 0)	1.04	(0.67, $\rightarrow$ 0)	1.04
-0.20	(0.75, $\rightarrow$ 0)	1.04	(0.83, $\rightarrow$ 0)	1.04
0.00	(1.00, $\rightarrow$ 0)	1.04	(1.00, $\rightarrow$ 0)	1.04
0.20	(1.00, $\rightarrow$ 0)	1.08	(1.00, $\rightarrow$ 0)	1.08
0.40	(1.00, $\rightarrow$ 0)	1.20	(1.00, 0.21)	1.19
0.49	(1.00, $\rightarrow$ 0)	1.28	(1.00, 0.41)	1.25
0.60	(1.00, $\rightarrow$ 0)	1.40	(1.00, 0.66)	1.31
0.79	(1.00, $\rightarrow$ 0)	1.66	(1.00, 1.00)	1.43
1.00	-	-	(1.00, 1.00)	1.68
1.19	-	-	(1.00, 1.00)	2.06
	$\theta_2 = -0.5$		$\theta_2 = 0.5$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-1.49	-	-	(0.01, $\rightarrow$ 0)	1.25
-1.40	-	-	(0.07, $\rightarrow$ 0)	1.25
-1.19	-	-	(0.21, $\rightarrow$ 0)	1.25
-1.00	-	-	(0.33, $\rightarrow$ 0)	1.25
-0.79	-	-	(0.47, $\rightarrow$ 0)	1.25
-0.60	-	-	(0.60, $\rightarrow$ 0)	1.25
-0.49	(0.02, $\rightarrow$ 0)	1.25	(0.67, $\rightarrow$ 0)	1.25
-0.40	(0.20, $\rightarrow$ 0)	1.25	(0.73, $\rightarrow$ 0)	1.25
-0.20	(0.60, $\rightarrow$ 0)	1.25	(0.87, $\rightarrow$ 0)	1.25
0.00	(1.00, $\rightarrow$ 0)	1.25	(1.00, $\rightarrow$ 0)	1.25
0.20	(1.00, $\rightarrow$ 0)	1.29	(1.00, 0.16)	1.28
0.40	(1.00, $\rightarrow$ 0)	1.41	(1.00, 0.44)	1.31
0.49	(1.00, $\rightarrow$ 0)	1.49	(1.00, 0.57)	1.33
0.60	-	-	(1.00, 0.73)	1.35
0.79	-	-	(1.00, 1.00)	1.38
1.00	-	-	(1.00, 1.00)	1.50
1.19	-	-	(1.00, 1.00)	1.76

3.3.3 N_t is an ARIMA$(0,0,q)$ process

N_t is an ARIMA$(0,0,q)$ process

$$N_t = \sum_{i=0}^{q} \theta_i \epsilon_{t-i}. \tag{3.68}$$

Let $\theta_{-2} = \theta_{-1} = \theta_{q+1} = \theta_{q+2} = 0$, then

$$(1-B)^2 N_t = \sum_{j=0}^{q+2} (\theta_j - 2\theta_{j-1} + \theta_{j-2}) \epsilon_{t-j}. \tag{3.69}$$

1). $q = 0$, N_t is a white noise.

Figure 3.19 shows that the MSE is a monotone decreasing function of both α_1 and α_2, and the minimum MSE occurs as $\alpha_1 \to 0$ and $\alpha_2 \to 0$.

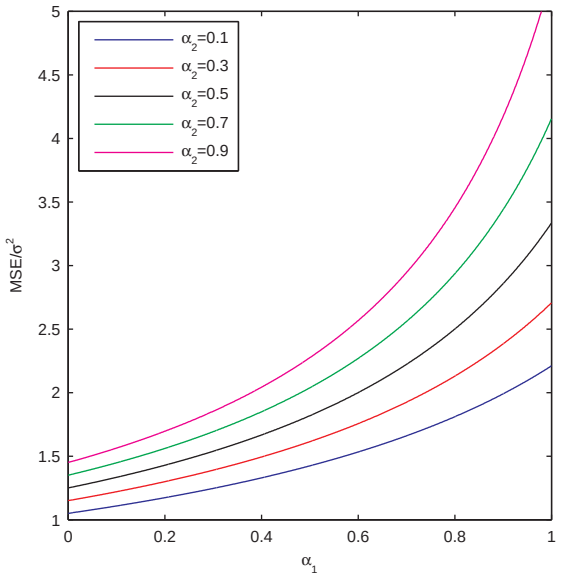


Figure 3.19: Holt's Method – MSE/σ^2 as a function of α_1, N_t is a white noise

2). $q = 1$, N_t is an MA(1).

Figure 3.20 and Table 3.14 show that

- The minimum MSE depends only on $|\theta_1|$, the absolute value of θ_1, and is free of the sign of θ_1. Moreover, the minimum MSE increases as $|\theta_1|$ increases.

– The MSE is a monotone decreasing function of both α_1 and α_2, and reaches minimum as $\alpha_1 \to 0$ and $\alpha_2 \to 0$ for any value of θ_1.

– The larger the θ_1, the less critical the choice of α_1. When θ_1, say, is greater than 0.5, any choice in the interval (0,1] for α_1 will be equally good.

In summary, Holt's method performs well for an MA(1) only when $|\theta_1|$ is small, and small α_1 and α_2 are preferred.

Table 3.14: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is a MA(1)

θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.95	($\to 0$, $\to 0$)	1.90	0.95	($\to 0$, $\to 0$)	1.90
-0.90	($\to 0$, $\to 0$)	1.81	0.90	($\to 0$, $\to 0$)	1.81
-0.80	($\to 0$, $\to 0$)	1.64	0.80	($\to 0$, $\to 0$)	1.64
-0.70	($\to 0$, $\to 0$)	1.49	0.70	($\to 0$, $\to 0$)	1.49
-0.60	($\to 0$, $\to 0$)	1.36	0.60	($\to 0$, $\to 0$)	1.36
-0.50	($\to 0$, $\to 0$)	1.25	0.50	($\to 0$, $\to 0$)	1.25
-0.40	($\to 0$, $\to 0$)	1.16	0.40	($\to 0$, $\to 0$)	1.16
-0.30	($\to 0$, $\to 0$)	1.09	0.30	($\to 0$, $\to 0$)	1.09
-0.20	($\to 0$, $\to 0$)	1.04	0.20	($\to 0$, $\to 0$)	1.04
-0.10	($\to 0$, $\to 0$)	1.01	0.10	($\to 0$, $\to 0$)	1.01

3). $q = 2$, N_t is an MA(2).

Figure 3.21 and Table 3.15 show that Holt's method performs well only when both $|theta_1|$ and $|\theta_2|$ are small. The optimal value of both α_1 and α_2 are small.

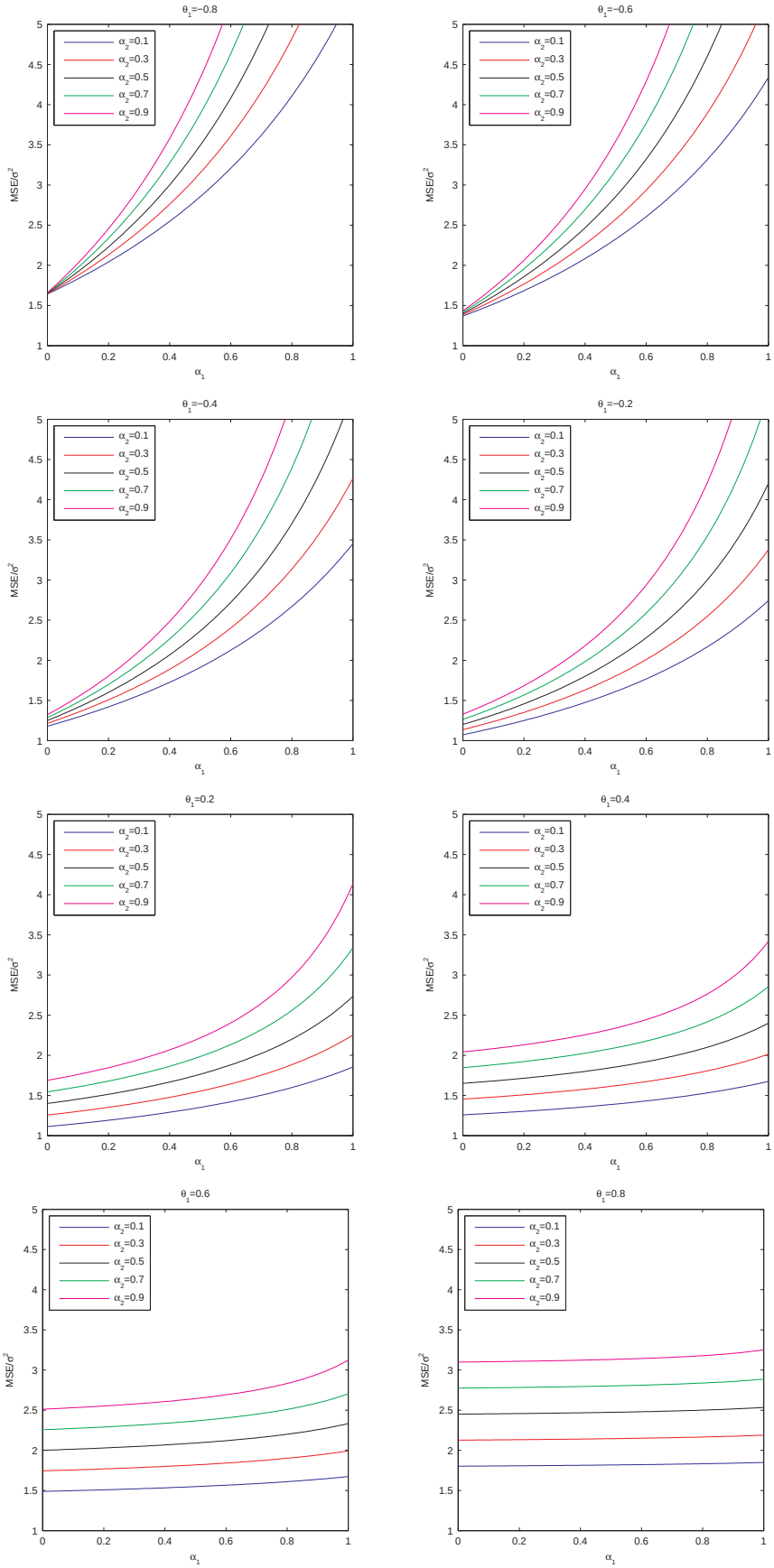


Figure 3.20: Holt's Method – MSE/σ^2 as a function of α_1, N_t is a MA(1)

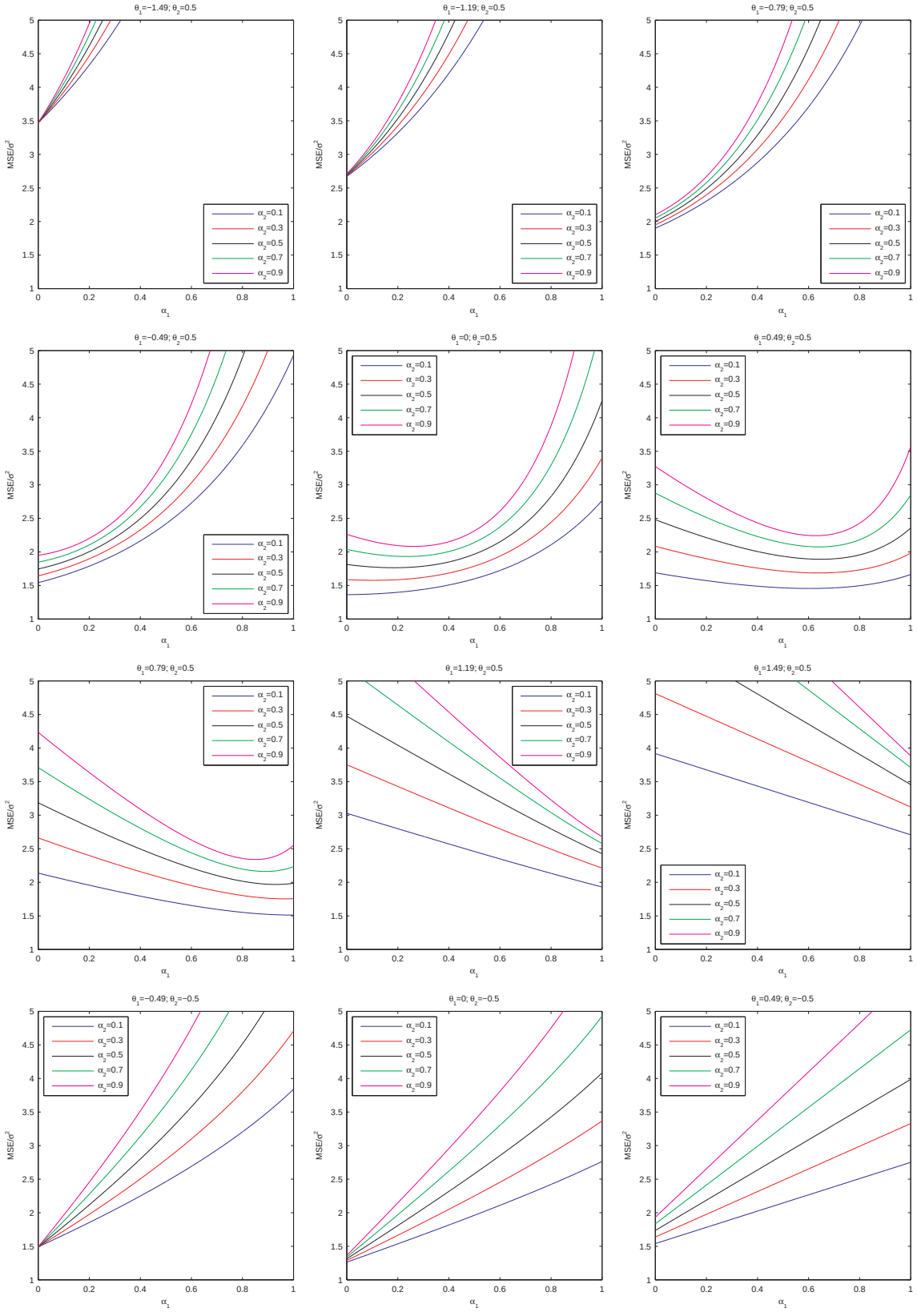


Figure 3.21: Holt's Method – MSE/σ^2 as a function of α_1, N_t is a MA(2)

Table 3.15: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is a MA(2)

	$\theta_2 = -0.2$		$\theta_2 = 0.2$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-1.19	-	-	$(\rightarrow 0, \rightarrow 0)$	2.46
-1.00	-	-	$(\rightarrow 0, \rightarrow 0)$	2.04
-0.79	$(\rightarrow 0, \rightarrow 0)$	1.66	$(\rightarrow 0, \rightarrow 0)$	1.66
-0.60	$(\rightarrow 0, \rightarrow 0)$	1.40	$(\rightarrow 0, \rightarrow 0)$	1.40
-0.49	$(\rightarrow 0, \rightarrow 0)$	1.28	$(\rightarrow 0, \rightarrow 0)$	1.28
-0.40	$(\rightarrow 0, \rightarrow 0)$	1.20	$(\rightarrow 0, \rightarrow 0)$	1.20
-0.20	$(\rightarrow 0, \rightarrow 0)$	1.08	$(\rightarrow 0, \rightarrow 0)$	1.08
0.00	$(\rightarrow 0, \rightarrow 0)$	1.04	$(\rightarrow 0, \rightarrow 0)$	1.04
0.20	$(\rightarrow 0, \rightarrow 0)$	1.08	$(\rightarrow 0, \rightarrow 0)$	1.08
0.40	$(\rightarrow 0, \rightarrow 0)$	1.20	$(0.21, \rightarrow 0)$	1.19
0.49	$(\rightarrow 0, \rightarrow 0)$	1.28	$(0.41, \rightarrow 0)$	1.25
0.60	$(\rightarrow 0, \rightarrow 0)$	1.40	$(0.66, \rightarrow 0)$	1.31
0.79	$(\rightarrow 0, \rightarrow 0)$	1.66	$(1.00, \rightarrow 0)$	1.43
1.00	-	-	$(1.00, \rightarrow 0)$	1.68
1.19	-	-	$(1.00, \rightarrow 0)$	2.06
	$\theta_2 = -0.5$		$\theta_2 = 0.5$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-1.49	-	-	$(\rightarrow 0, \rightarrow 0)$	3.47
-1.40	-	-	$(\rightarrow 0, \rightarrow 0)$	3.21
-1.19	-	-	$(\rightarrow 0, \rightarrow 0)$	2.67
-1.00	-	-	$(\rightarrow 0, \rightarrow 0)$	2.25
-0.79	-	-	$(\rightarrow 0, \rightarrow 0)$	1.87
-0.60	-	-	$(\rightarrow 0, \rightarrow 0)$	1.61
-0.49	$(\rightarrow 0, \rightarrow 0)$	1.49	$(\rightarrow 0, \rightarrow 0)$	1.49
-0.40	$(\rightarrow 0, \rightarrow 0)$	1.41	$(\rightarrow 0, \rightarrow 0)$	1.41
-0.20	$(\rightarrow 0, \rightarrow 0)$	1.29	$(\rightarrow 0, \rightarrow 0)$	1.29
0.00	$(\rightarrow 0, \rightarrow 0)$	1.25	$(\rightarrow 0, \rightarrow 0)$	1.25
0.20	$(\rightarrow 0, \rightarrow 0)$	1.29	$(0.16, \rightarrow 0)$	1.28
0.40	$(\rightarrow 0, \rightarrow 0)$	1.41	$(0.44, \rightarrow 0)$	1.31
0.49	$(\rightarrow 0, \rightarrow 0)$	1.49	$(0.57, \rightarrow 0)$	1.33
0.60	-	-	$(0.73, \rightarrow 0)$	1.35
0.79	-	-	$(1.00, \rightarrow 0)$	1.38
1.00	-	-	$(1.00, \rightarrow 0)$	1.50
1.19	-	-	$(1.00, \rightarrow 0)$	1.76

3.3.4 N_t is an ARIMA(1,d,0) process

N_t is an ARIMA(1,d,0) process

$$(1-B)^d N_t = (1-\phi_1 B)^{-1}\epsilon_t. \tag{3.70}$$

1). $d = 2$, N_t is an ARIMA(1,2,0).

$$(1-B)^2 N_t = (1-\phi_1 B)^{-1}\epsilon_t = \sum_{i=0}^{\infty} \phi_1^i \epsilon_{t-i}. \tag{3.71}$$

Figure 3.22 and Table 3.16 show that

- Holt's method performs better when $-1 < \phi_1 < 0$ than when $0 < \phi_1 < 1$. When $-1 < \phi_1 < 0$, the minimum MSE increases as $|\phi_1|$ increases, and the larger the $|\phi_1|$, the faster the minimum MSE grows. Same holds for $0 < \phi_1 < 1$.
- The optimal values of α_1 and α_2 are typically greater than 0.5 except when ϕ_1 is very close to -1.

In summary, Holt's method performs well for an ARIMA(1,2,0) when either $-1 < \phi_1 < 0$ or $|\phi_1|$ is small, and large α_1 and α_2 are perferred.

Table 3.16: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARIMA(1,2,0)

ϕ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	ϕ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.95	(0.45, 1.00)	2.33	0.95	(1.00, 1.00)	10.26
-0.90	(0.52, 1.00)	1.61	0.90	(1.00, 1.00)	5.26
-0.80	(0.61, 1.00)	1.22	0.80	(1.00, 1.00)	2.78
-0.70	(0.67, 0.99)	1.09	0.70	(1.00, 1.00)	1.96
-0.60	(0.74, 0.92)	1.04	0.60	(1.00, 1.00)	1.56
-0.50	(0.80, 0.89)	1.01	0.50	(1.00, 1.00)	1.33
-0.40	(0.86, 0.87)	1.00	0.40	(1.00, 1.00)	1.19
-0.30	(0.92, 0.85)	1.00	0.30	(1.00, 1.00)	1.10
-0.20	(0.96, 0.88)	1.00	0.20	(1.00, 1.00)	1.04
-0.10	(0.99, 0.92)	1.00	0.10	(1.00, 1.00)	1.01

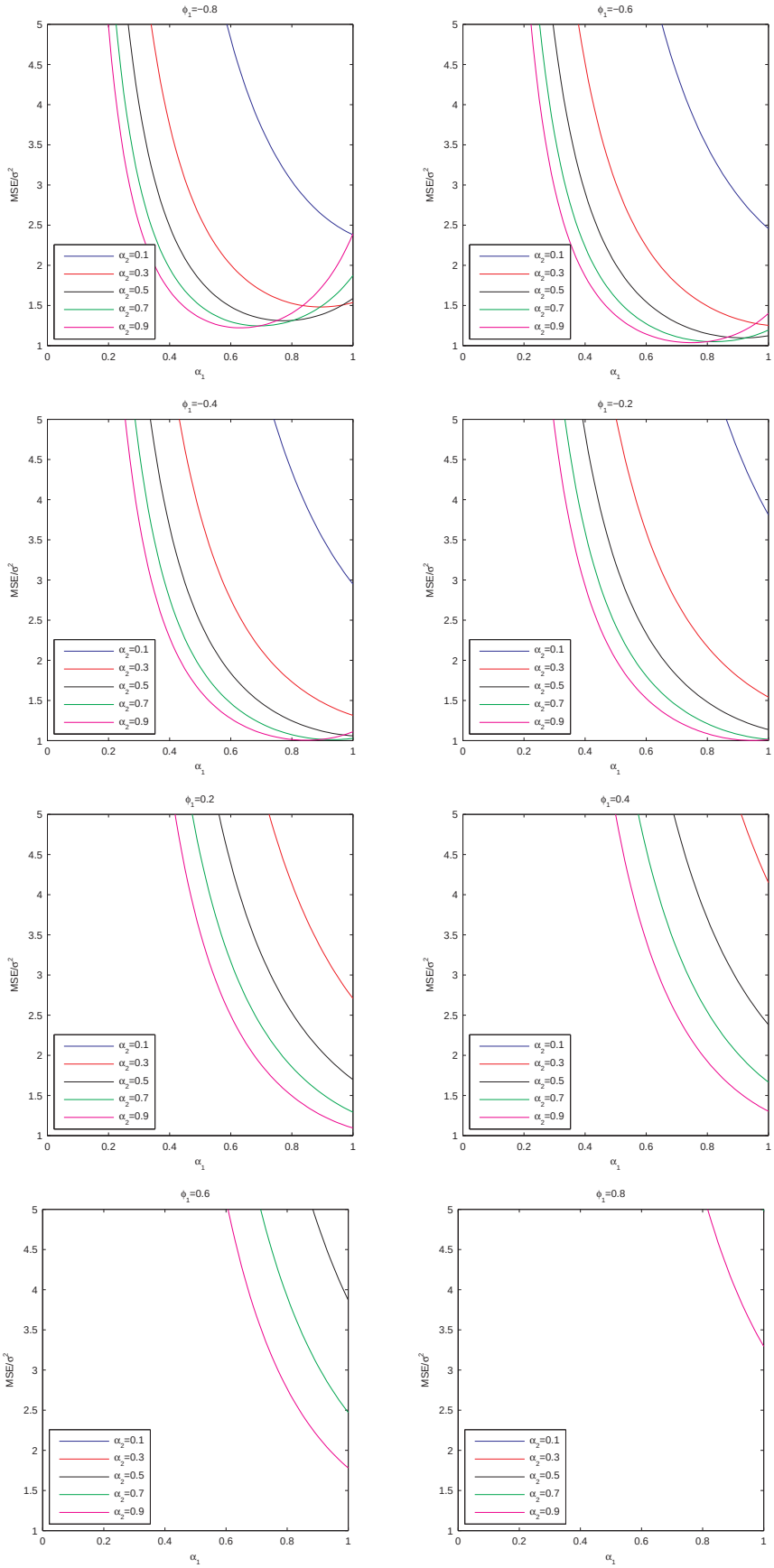


Figure 3.22: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(1,2,0)

2). $d = 1$, N_t is an ARIMA(1,1,0).

$$(1 - B)N_t = (1 - \phi_1 B)^{-1}\epsilon_t. \tag{3.72}$$

Then

$$(1 - B)^2 N_t = (1 - B)(1 - \phi_1 B)^{-1}\epsilon_t = \epsilon_t + (\phi_1 - 1)\sum_{i=1}^{\infty} \phi_1^{i-1}\epsilon_{t-i}. \tag{3.73}$$

Figure 3.23 and Table 3.17 show that

- Holt's method performs well for an ARIMA(1,1,0) disturbance except when ϕ_1 is close to -1. When $-1 < \phi_1 < 0$, the minimum MSE increases as ϕ_1 decreases. When $0 < \phi_1 < 1$, the worst performance occurs at $\phi_1 = 0.5$, and the further away from 0.5 the ϕ_1, the smaller the minimum MSE.
- When $-1 < \phi_1 < 0$, the optimal value of α_1 increases as ϕ_1 increases, and overestimation of α_1 is less serious than the equivalent underestimation. When $0 < \phi_1 < 1$, the optimal value of α_1 stays at 1.
- When $-1 < \phi_1 \leq 0.3$, the optimal value of α_2 stays at an extremely small value ($\rightarrow 0$); when $0.3 < \phi_1 < 1$, the optimal value of α_2 increases as ϕ_1 increases; and the choice of α_2 is not critical when $0 < \phi_1 < 1$ and $|\phi_1|$ is small.

In summary, Holt's method performs well for an ARIMA(1,1,0), and a large α_1 is preferred.

3). $d = 0$, N_t is an AR(1).

$$N_t = (1 - \phi_1 B)^{-1}\epsilon_t. \tag{3.74}$$

Then

$$\begin{aligned}(1 - B)^2 N_t &= (1 - B)^2(1 - \phi_1 B)^{-1}\epsilon_t \\ &= \epsilon_t + (\phi_1 - 2)\epsilon_{t-1} + (\phi_1 - 1)^2 \sum_{i=2}^{\infty} \phi_1^{i-2}\epsilon_{t-i}. \end{aligned} \tag{3.75}$$

Figure 3.24 and Table 3.18 show that

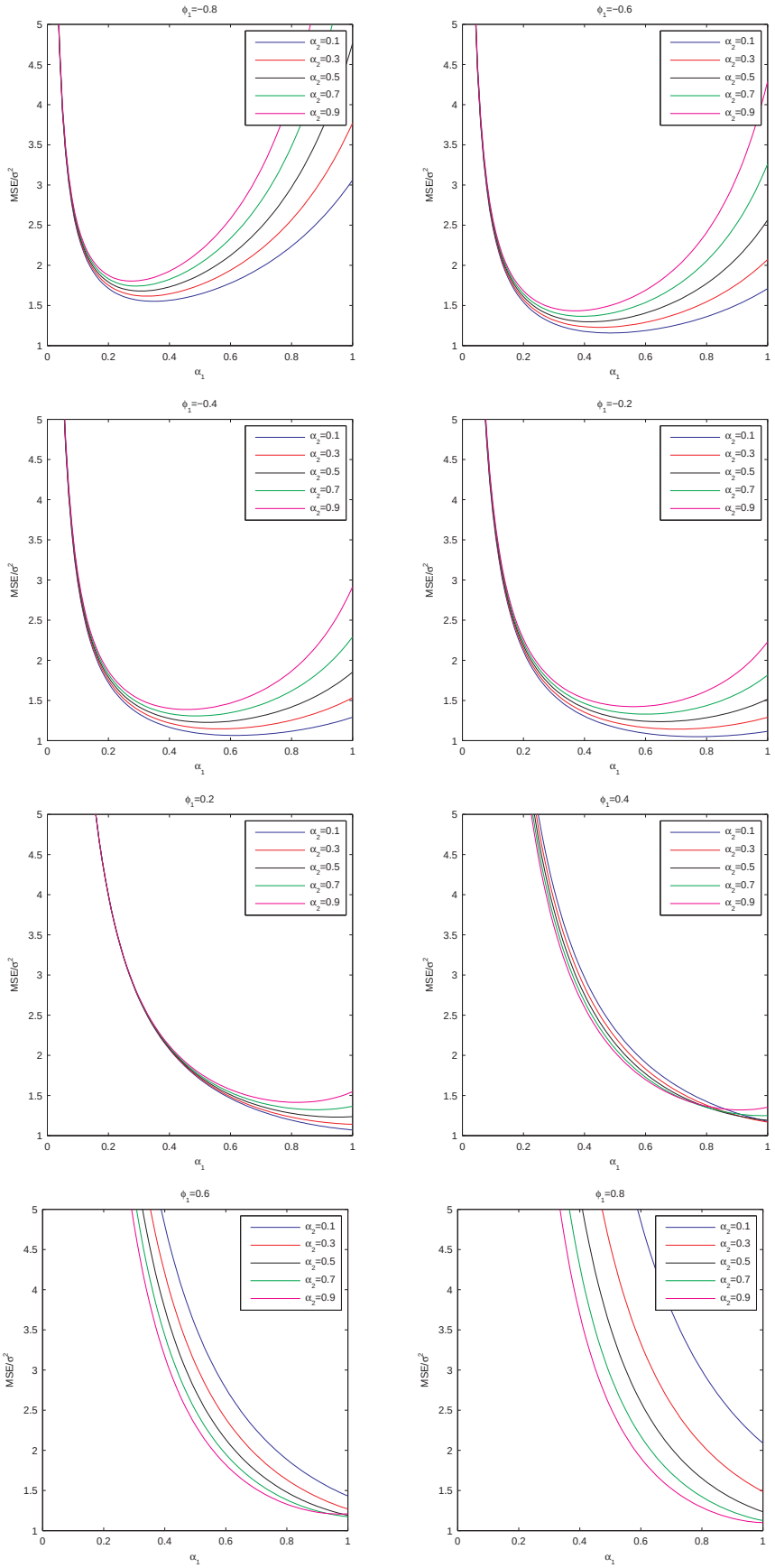


Figure 3.23: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(1,1,0)

Table 3.17: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARIMA(1,1,0)

ϕ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	ϕ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.95	(0.19, $\rightarrow$ 0)	3.88	0.95	(1.00, 0.97)	1.02
-0.90	(0.26, $\rightarrow$ 0)	2.33	0.90	(1.00, 0.94)	1.05
-0.80	(0.36, $\rightarrow$ 0)	1.52	0.80	(1.00, 0.87)	1.10
-0.70	(0.44, $\rightarrow$ 0)	1.25	0.70	(1.00, 0.79)	1.14
-0.60	(0.50, $\rightarrow$ 0)	1.12	0.60	(1.00, 0.67)	1.17
-0.50	(0.57, $\rightarrow$ 0)	1.06	0.50	(1.00, 0.50)	1.19
-0.40	(0.64, $\rightarrow$ 0)	1.02	0.40	(1.00, 0.25)	1.17
-0.30	(0.72, $\rightarrow$ 0)	1.01	0.30	(1.00, $\rightarrow$ 0)	1.10
-0.20	(0.81, $\rightarrow$ 0)	1.00	0.20	(1.00, $\rightarrow$ 0)	1.04
-0.10	(0.90, $\rightarrow$ 0)	1.00	0.10	(1.00, $\rightarrow$ 0)	1.01

- Holt's method performs well for an AR(1) disturbance except when ϕ_1 is close to -1. When $-1 < \phi_1 < 0$, the minimum MSE increases as ϕ_1 decreases. When $0 < \phi_1 < 1$, the worst performance occurs at $\phi_1 = 0.5$, and the further away from 0.5 the ϕ_1, the smaller the minimum MSE.
- When $-1 < \phi_1 \leq 0.3$, the optimal value of α_1 stays at an extremely small value ($\rightarrow 0$); when $0.3 < \phi_1 < 1$, the optimal value of α_1 increases as ϕ_1 increases; and the choice of α_1 is not critical when ϕ_1 is close to 0.5.
- The optimal value of α_2 stays at an extremely small value ($\rightarrow 0$).

In summary, Holt's method performs well for an AR(1) except when ϕ_1 is close to -1, and a small α_2 is preferred.

Comparing the results for ARIMA(1,1,0) and AR(1) from SES and Holt's method respectively (Table 3.5 vs. Table 3.17, Table 3.6 vs. Table 3.18) reveals that Holt's method with $\alpha_2 \rightarrow 0$ for model (3.45) produces exactly the same results as these by SES for model (3.2) (see explanation given for ARIMA$(0, 1, q)$).

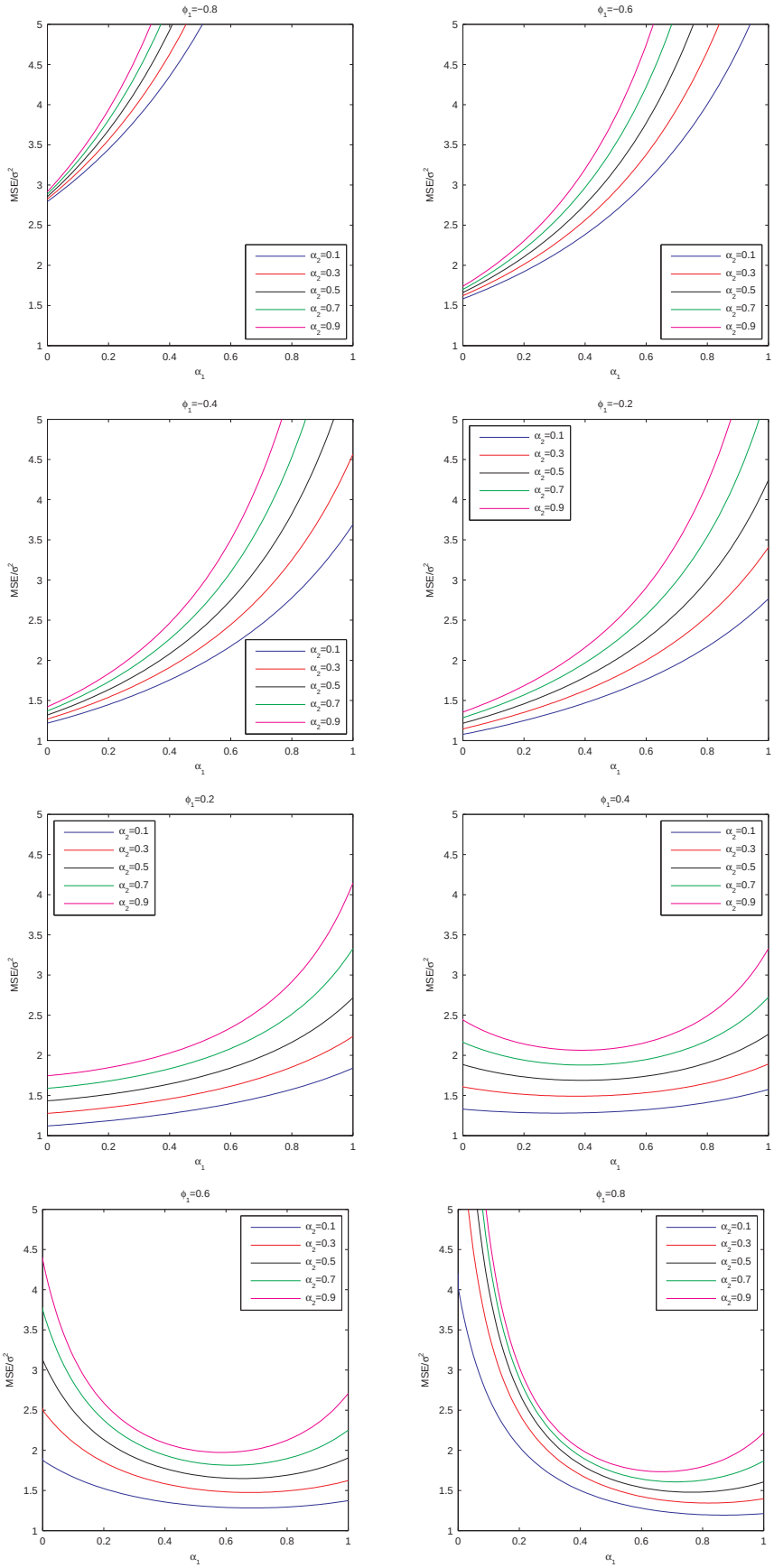


Figure 3.24: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an AR(1)

Table 3.18: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an AR(1)

ϕ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	ϕ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.95	($\rightarrow$ 0, 0.01)	10.26	0.95	(0.97, $\rightarrow$ 0)	1.02
-0.90	($\rightarrow$ 0, 0.01)	5.26	0.90	(0.94, $\rightarrow$ 0)	1.05
-0.80	($\rightarrow$ 0, $\rightarrow$ 0)	2.78	0.80	(0.88, $\rightarrow$ 0)	1.10
-0.70	($\rightarrow$ 0, $\rightarrow$ 0)	1.96	0.70	(0.79, $\rightarrow$ 0)	1.14
-0.60	($\rightarrow$ 0, $\rightarrow$ 0)	1.56	0.60	(0.67, $\rightarrow$ 0)	1.17
-0.50	($\rightarrow$ 0, $\rightarrow$ 0)	1.33	0.50	(0.50, $\rightarrow$ 0)	1.19
-0.40	($\rightarrow$ 0, $\rightarrow$ 0)	1.19	0.40	(0.25, $\rightarrow$ 0)	1.17
-0.30	($\rightarrow$ 0, $\rightarrow$ 0)	1.10	0.30	($\rightarrow$ 0, $\rightarrow$ 0)	1.10
-0.20	($\rightarrow$ 0, $\rightarrow$ 0)	1.04	0.20	($\rightarrow$ 0, $\rightarrow$ 0)	1.04
-0.10	($\rightarrow$ 0, $\rightarrow$ 0)	1.01	0.10	($\rightarrow$ 0, $\rightarrow$ 0)	1.01

3.3.5 N_t is an ARIMA(1,d,1) process

N_t is an ARIMA(1,d,1) process

$$(1-B)^d N_t = (1-\phi_1 B)^{-1}(\epsilon_t + \theta_1 \epsilon_{t-1}). \tag{3.76}$$

1). $d = 2$, N_t is an ARIMA(1,2,1).

$$(1-B)^2 N_t = (1-\phi_1 B)^{-1}(\epsilon_t + \theta_1 \epsilon_{t-1}) = \epsilon_t + (\phi_1 + \theta_1)\sum_{i=1}^{\infty} \phi_1^{i-1}\epsilon_{t-i}. \tag{3.77}$$

Figures 3.25 and 3.26 and Table 3.19 show that

- Holt's method performs well under two situations: i) $-1 < \theta_1 < 0$ and $|\phi_1|$ is small , and ii) $0 < \theta_1 < 1$, $\phi_1 < 0$ and $|\theta_1|$ and $|\phi_1|$ are close.
- α_1 increases as θ_1 and/or ϕ_1 increase, and α_2 increases as θ_1 increases.
- Overestimation of α_1 is less serious than the equivalent underestimation.

2). $d = 1$, N_t is an ARIMA(1,1,1)

$$(1-B)N_t = (1-\phi_1 B)^{-1}(\epsilon_t + \theta_1 \epsilon_{t-1}). \tag{3.78}$$

Then

$$(1-B)^2 N_t = \epsilon_t + (\phi_1 + \theta_1 - 1)\epsilon_{t-1} + (\phi_1 + \theta_1)(\phi_1 - 1)\sum_{i=2}^{\infty} \phi_1^{i-2}\epsilon_{t-i}. \tag{3.79}$$

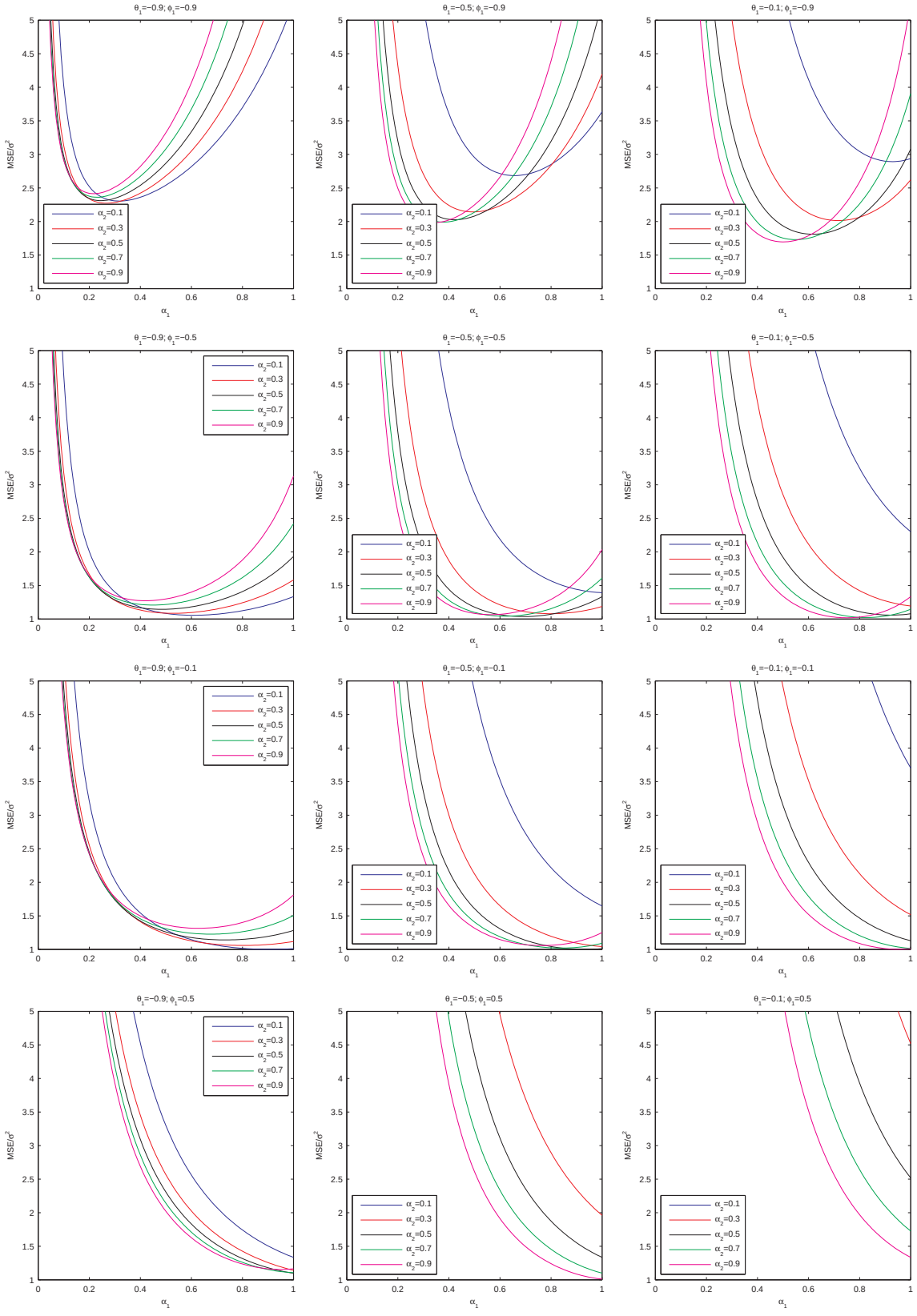


Figure 3.25: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(1,2,1)

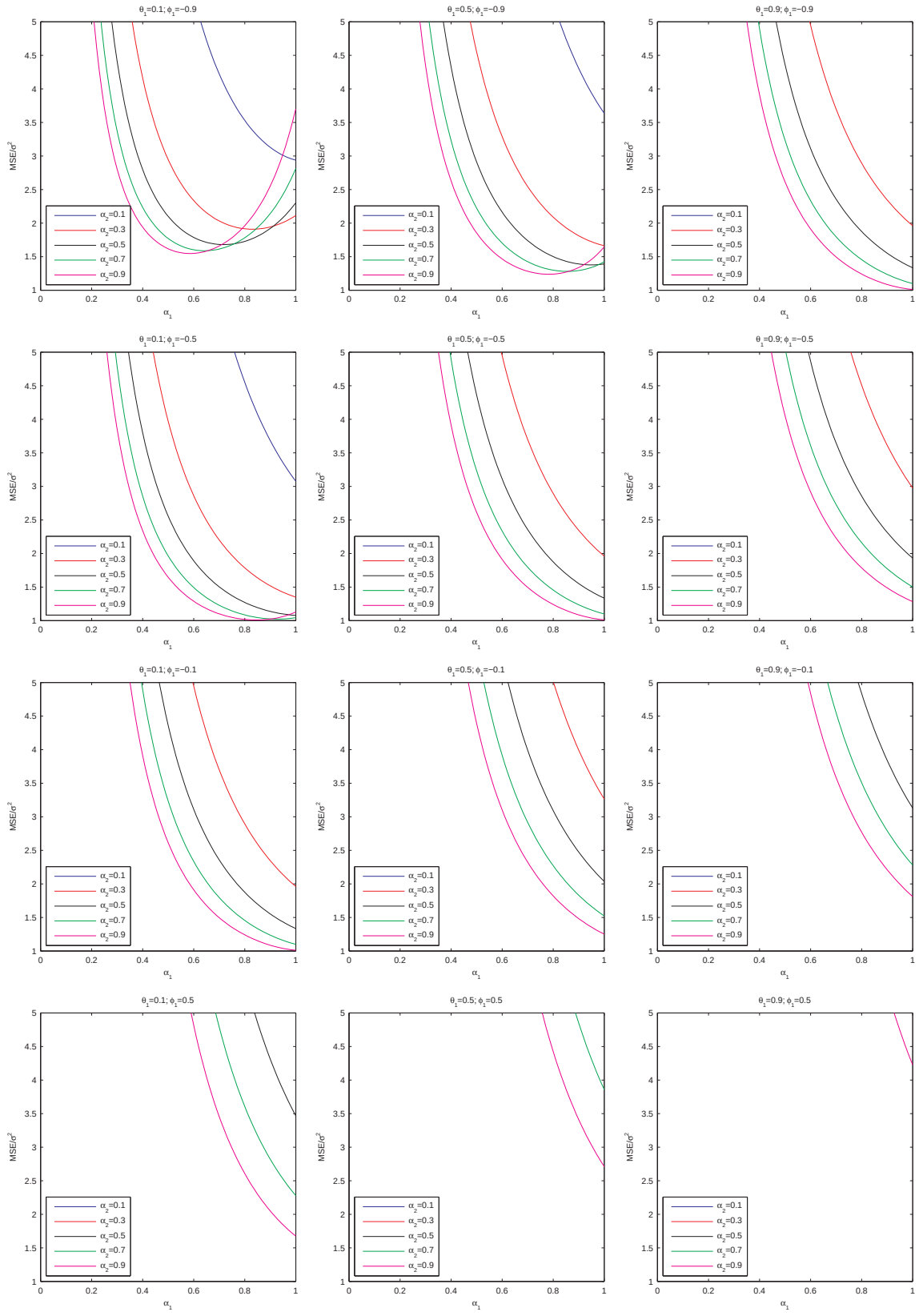


Figure 3.26: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(1,2,1)

Table 3.19: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARIMA(1,2,1)

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.90	(0.28, 0.22)	2.27	(0.59, 0.13)	1.05	(0.91, 0.10)	1.00
-0.70	(0.31, 0.59)	2.13	(0.63, 0.36)	1.04	(0.92, 0.30)	1.00
-0.50	(0.35, 0.86)	1.98	(0.67, 0.55)	1.03	(0.94, 0.49)	1.00
-0.30	(0.40, 1.00)	1.84	(0.72, 0.71)	1.03	(0.96, 0.67)	1.00
-0.10	(0.48, 1.00)	1.69	(0.77, 0.84)	1.02	(0.98, 0.84)	1.00

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.90	(1.00, 1.00)	1.00	(1.00, 0.56)	1.09	(1.00, 0.12)	1.01
-0.70	(1.00, 1.00)	1.21	(1.00, 0.79)	1.02	(1.00, 0.36)	1.00
-0.50	(1.00, 1.00)	1.84	(1.00, 1.00)	1.00	(1.00, 0.58)	1.00
-0.30	(1.00, 1.00)	2.89	(1.00, 1.00)	1.05	(1.00, 0.80)	1.00
-0.10	(1.00, 1.00)	4.37	(1.00, 1.00)	1.21	(1.00, 1.00)	1.00

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
0.90	(1.00, 1.00)	1.00	(1.00, 1.00)	1.21	(1.00, 1.00)	1.65
0.70	(0.87, 1.00)	1.08	(1.00, 1.00)	1.05	(1.00, 1.00)	1.36
0.50	(0.75, 1.00)	1.23	(1.00, 1.00)	1.00	(1.00, 1.00)	1.16
0.30	(0.65, 1.00)	1.38	(0.91, 0.98)	1.00	(1.00, 1.00)	1.04
0.10	(0.56, 1.00)	1.54	(0.84, 0.92)	1.01	(1.00, 1.00)	1.00

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
0.90	(1.00, 1.00)	18.05	(1.00, 1.00)	3.61	(1.00, 1.00)	2.01
0.70	(1.00, 1.00)	14.47	(1.00, 1.00)	2.92	(1.00, 1.00)	1.65
0.50	(1.00, 1.00)	11.32	(1.00, 1.00)	2.33	(1.00, 1.00)	1.36
0.30	(1.00, 1.00)	8.58	(1.00, 1.00)	1.85	(1.00, 1.00)	1.16
0.10	(1.00, 1.00)	6.26	(1.00, 1.00)	1.48	(1.00, 1.00)	1.04

Figures 3.27 and 3.28 and Table 3.20 show that

- Holt's method performs well $\theta_1\phi_1 < 0$.
- α_1 increases as θ_1 and/or ϕ_1 increase.
- Overestimation of α_1 is less serious than the equivalent underestimation, and, when $0 < \theta_1 < 1$ and $|\phi_1|$ is small, the choice of α_2 is not critical.

3). $d = 0$, N_t is an ARMA(1,1)

$$N_t = (1 - \phi_1 B)^{-1}(\epsilon_t + \theta_1\epsilon_{t-1}). \tag{3.80}$$

Then

$$\begin{aligned}(1 - B)^2 N_t &= \epsilon_t + (\phi_1 + \theta_1 - 2)\epsilon_{t-1} + [(\phi_1 + \theta_1)(\phi_1 - 2) + 1]\epsilon_{t-2} \\ &\quad + (\phi_1 + \theta_1)(\phi_1 - 1)^2 \sum_{i=3}^{\infty} \phi_1^{i-3}\epsilon_{t-i}. \end{aligned} \tag{3.81}$$

Figures 3.29 and 3.30 and Table 3.21 show that

- Holt's method performs well when $\theta_1\phi_1 < 0$ and $|\theta_1|$ and $|\phi_1|$ are close.
- The optimal value of α_1, starting from an small value (≈ 0), becomes not critical and then increases as θ_1 and/or ϕ_1 increases.
- The optimal value of α_2 stays at an small value (≈ 0) regardless of what value θ_1 or ϕ_1 takes.

3.3.6 Summary

Based on the results above on the performance of Holt's method for different types of ARIMA time series, the following conclusions can be drawn:

- N_t is an ARIMA$(0, 2, q)$ with $0 \leq q \leq 2$. $\boldsymbol{\alpha}_{opt} = (1, 1)^T$ when $\theta_1 > 0$, $\boldsymbol{\alpha}_{opt} < (1, 1)^T$ when $\theta_1 < 0$, and overestimation of either α_1 or α_2 is less serious than the equivalent underestimation. Holt's method performs well when $\theta_1 < 0$ and $0 \leq \theta_2 < 1$.

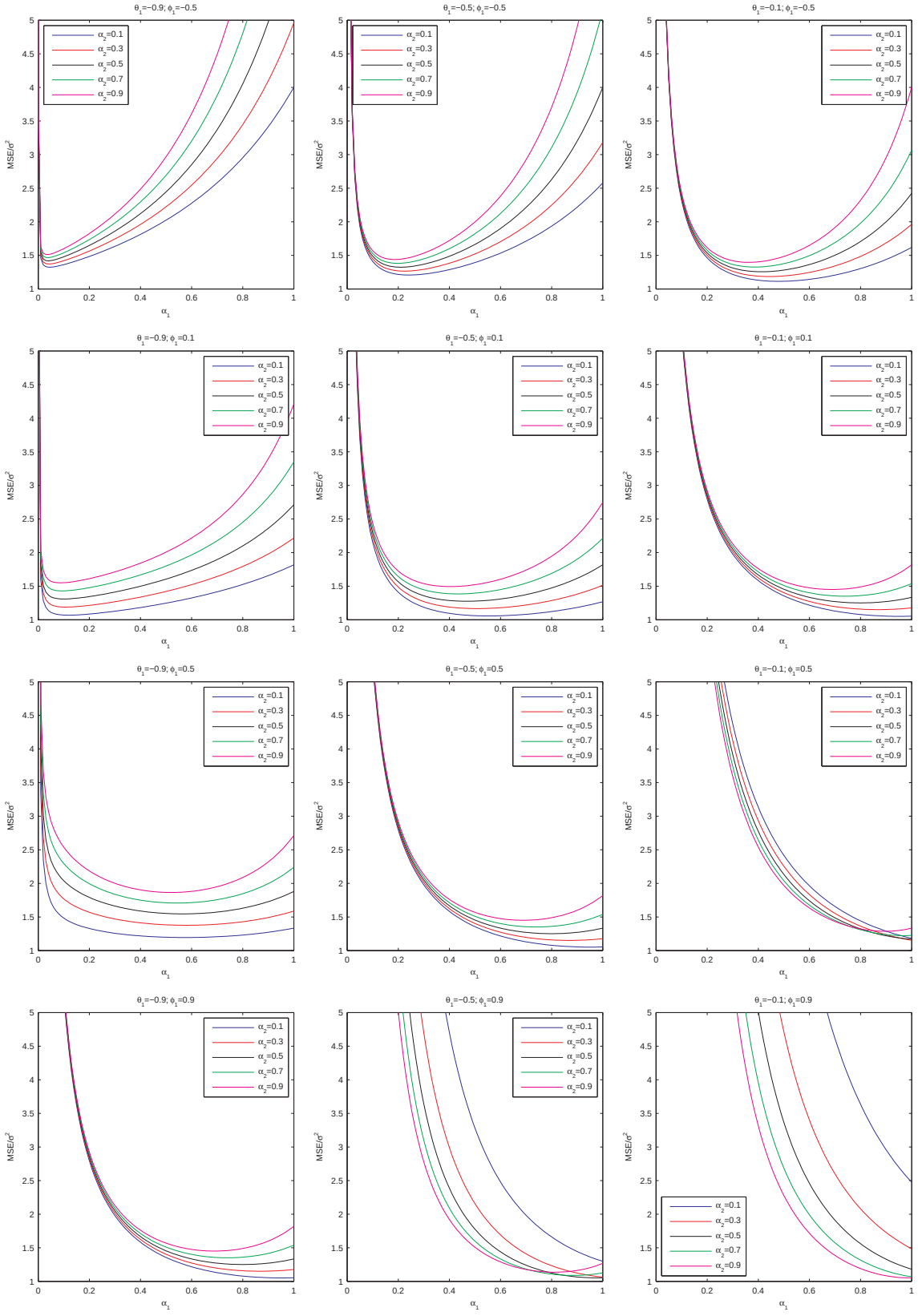


Figure 3.27: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(1,1,1)

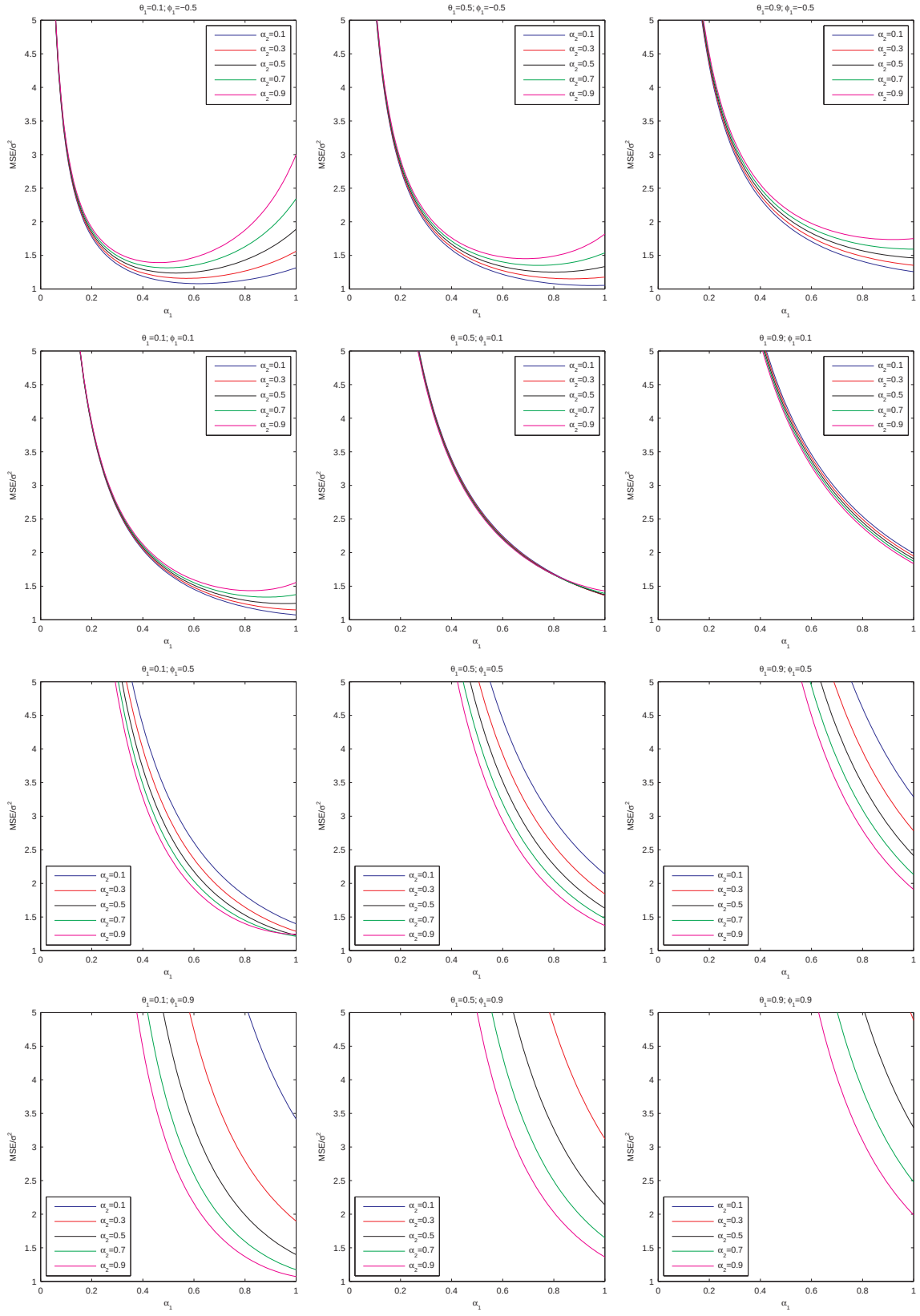


Figure 3.28: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARIMA(1,1,1)

Table 3.20: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARIMA(1,1,1)

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.90	(0.02, $\rightarrow$ 0)	4.91	(0.05, $\rightarrow$ 0)	1.30	(0.08, $\rightarrow$ 0)	1.01
-0.70	(0.06, $\rightarrow$ 0)	4.25	(0.14, $\rightarrow$ 0)	1.24	(0.26, $\rightarrow$ 0)	1.01
-0.50	(0.10, $\rightarrow$ 0)	3.63	(0.25, $\rightarrow$ 0)	1.18	(0.43, $\rightarrow$ 0)	1.00
-0.30	(0.16, $\rightarrow$ 0)	3.07	(0.37, $\rightarrow$ 0)	1.12	(0.61, $\rightarrow$ 0)	1.00
-0.10	(0.22, $\rightarrow$ 0)	2.57	(0.50, $\rightarrow$ 0)	1.08	(0.80, $\rightarrow$ 0)	1.00

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.90	(1.00, $\rightarrow$ 0)	1.00	(0.56, $\rightarrow$ 0)	1.09	(0.12, $\rightarrow$ 0)	1.01
-0.70	(1.00, 0.19)	1.04	(0.79, $\rightarrow$ 0)	1.02	(0.36, $\rightarrow$ 0)	1.00
-0.50	(1.00, 0.41)	1.04	(1.00, $\rightarrow$ 0)	1.00	(0.58, $\rightarrow$ 0)	1.00
-0.30	(1.00, 0.63)	1.05	(1.00, $\rightarrow$ 0)	1.05	(0.80, $\rightarrow$ 0)	1.00
-0.10	(1.00, 0.84)	1.05	(1.00, 0.33)	1.15	(1.00, $\rightarrow$ 0)	1.00

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
0.90	(1.00, $\rightarrow$ 0)	1.00	(1.00, $\rightarrow$ 0)	1.21	(1.00, $\rightarrow$ 0)	1.65
0.70	(0.76, $\rightarrow$ 0)	1.12	(1.00, $\rightarrow$ 0)	1.05	(1.00, $\rightarrow$ 0)	1.36
0.50	(0.57, $\rightarrow$ 0)	1.38	(1.00, $\rightarrow$ 0)	1.00	(1.00, $\rightarrow$ 0)	1.16
0.30	(0.42, $\rightarrow$ 0)	1.71	(0.81, $\rightarrow$ 0)	1.01	(1.00, $\rightarrow$ 0)	1.04
0.10	(0.31, $\rightarrow$ 0)	2.11	(0.65, $\rightarrow$ 0)	1.04	(1.00, $\rightarrow$ 0)	1.00

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
0.90	(1.00, 1.00)	1.81	(1.00, 1.00)	1.81	(1.00, 1.00)	1.82
0.70	(1.00, 1.00)	1.49	(1.00, 1.00)	1.52	(1.00, 0.98)	1.56
0.50	(1.00, 1.00)	1.26	(1.00, 1.00)	1.33	(1.00, 0.17)	1.36
0.30	(1.00, 1.00)	1.12	(1.00, 0.98)	1.25	(1.00, $\rightarrow$ 0)	1.16
0.10	(1.00, 1.00)	1.05	(1.00, 0.67)	1.21	(1.00, $\rightarrow$ 0)	1.04

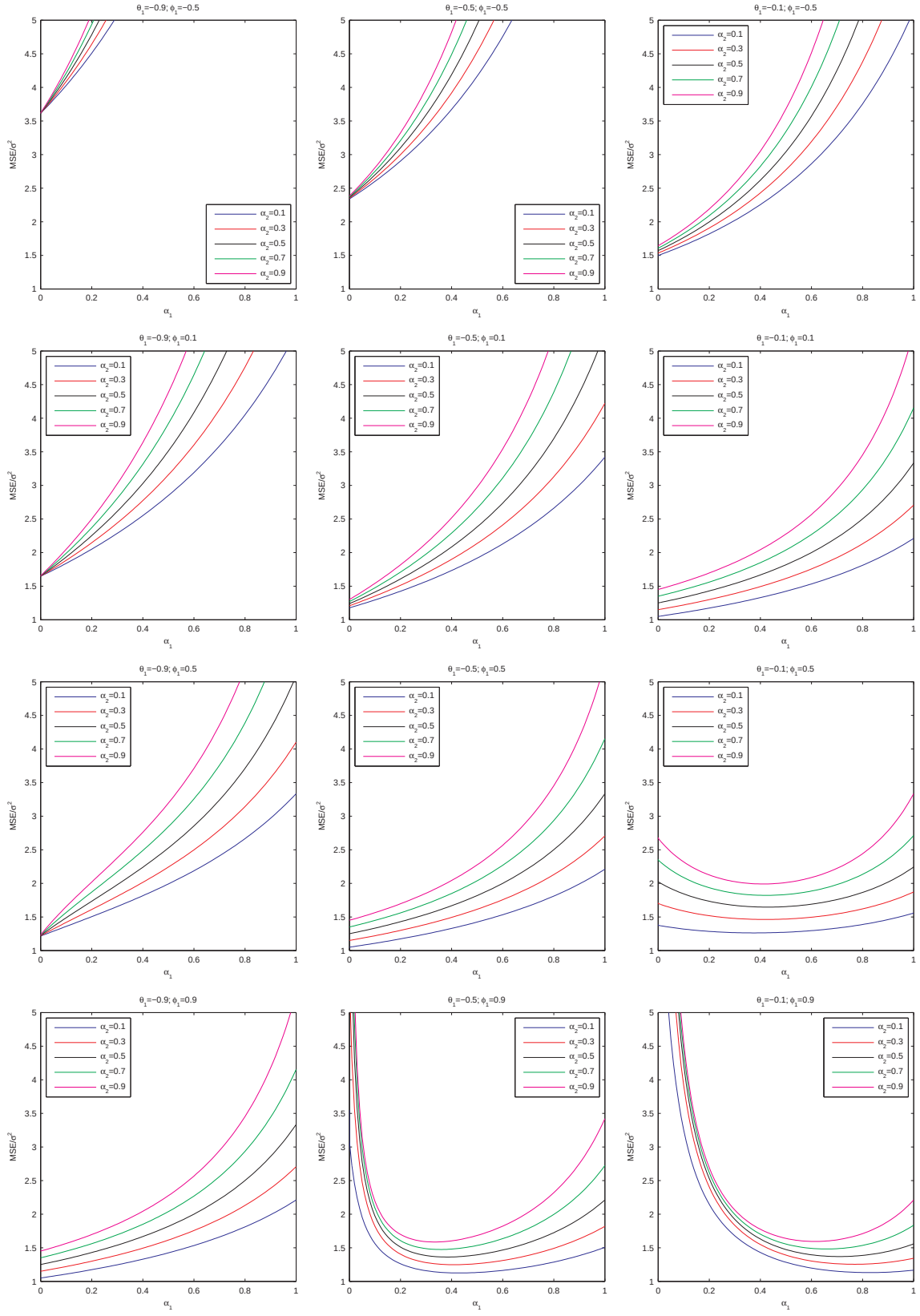


Figure 3.29: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARMA(1,1)

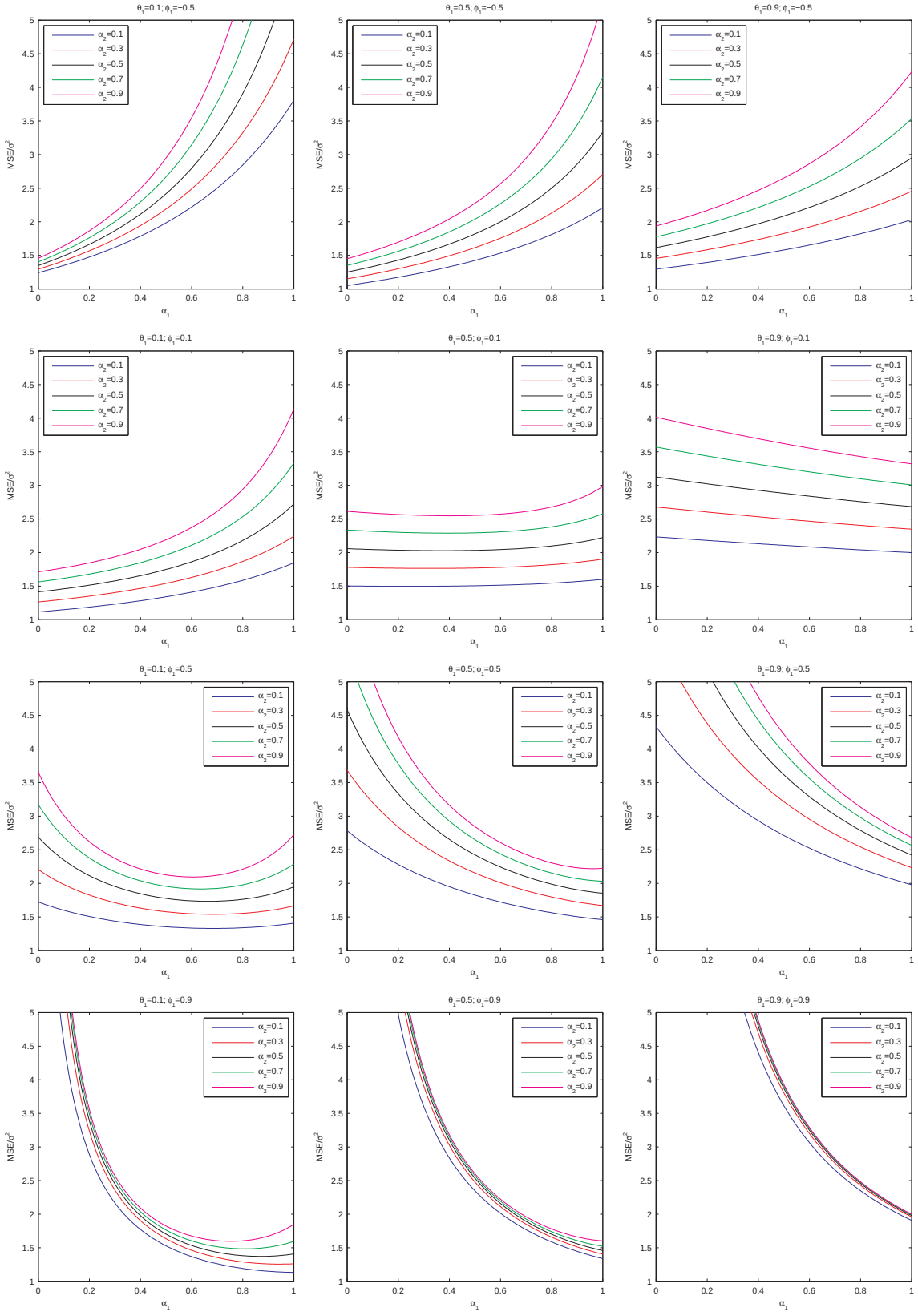


Figure 3.30: Holt's Method – MSE/σ^2 as a function of α_1, N_t is an ARMA(1,1)

Table 3.21: Holt's Method – Optimal $\boldsymbol{\alpha}$ and Minimum MSE/σ^2, N_t is an ARMA(1,1)

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.90	($\rightarrow$ 0, 0.09)	18.05	($\rightarrow$ 0, 0.03)	3.61	($\rightarrow$ 0, $\rightarrow$ 0)	2.01
-0.70	($\rightarrow$ 0, $\rightarrow$ 0)	14.47	($\rightarrow$ 0, 0.01)	2.92	($\rightarrow$ 0, $\rightarrow$ 0)	1.65
-0.50	($\rightarrow$ 0, 0.02)	11.32	($\rightarrow$ 0, $\rightarrow$ 0)	2.33	($\rightarrow$ 0, $\rightarrow$ 0)	1.36
-0.30	($\rightarrow$ 0, 0.01)	8.58	($\rightarrow$ 0, $\rightarrow$ 0)	1.85	($\rightarrow$ 0, $\rightarrow$ 0)	1.16
-0.10	($\rightarrow$ 0, $\rightarrow$ 0)	6.26	($\rightarrow$ 0, $\rightarrow$ 0)	1.48	($\rightarrow$ 0, $\rightarrow$ 0)	1.04

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
-0.90	($\rightarrow$ 0, $\rightarrow$ 0)	1.00	($\rightarrow$ 0, 0.03)	1.21	($\rightarrow$ 0, 0.01)	1.65
-0.70	(0.19, $\rightarrow$ 0)	1.04	($\rightarrow$ 0, 0.01)	1.05	($\rightarrow$ 0, $\rightarrow$ 0)	1.36
-0.50	(0.41, $\rightarrow$ 0)	1.04	($\rightarrow$ 0, $\rightarrow$ 0)	1.00	($\rightarrow$ 0, $\rightarrow$ 0)	1.16
-0.30	(0.63, $\rightarrow$ 0)	1.05	($\rightarrow$ 0, $\rightarrow$ 0)	1.05	($\rightarrow$ 0, $\rightarrow$ 0)	1.04
-0.10	(0.84, $\rightarrow$ 0)	1.05	(0.33, $\rightarrow$ 0)	1.15	($\rightarrow$ 0, $\rightarrow$ 0)	1.00

	$\phi_1 = -0.9$		$\phi_1 = -0.5$		$\phi_1 = -0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
0.90	($\rightarrow$ 0, $\rightarrow$ 0)	1.00	($\rightarrow$ 0, $\rightarrow$ 0)	1.21	($\rightarrow$ 0, $\rightarrow$ 0)	1.65
0.70	($\rightarrow$ 0, $\rightarrow$ 0)	1.21	($\rightarrow$ 0, $\rightarrow$ 0)	1.05	($\rightarrow$ 0, $\rightarrow$ 0)	1.36
0.50	($\rightarrow$ 0, $\rightarrow$ 0)	1.84	($\rightarrow$ 0, $\rightarrow$ 0)	1.00	($\rightarrow$ 0, $\rightarrow$ 0)	1.16
0.30	($\rightarrow$ 0, $\rightarrow$ 0)	2.89	($\rightarrow$ 0, $\rightarrow$ 0)	1.05	($\rightarrow$ 0, $\rightarrow$ 0)	1.04
0.10	($\rightarrow$ 0, $\rightarrow$ 0)	4.37	($\rightarrow$ 0, $\rightarrow$ 0)	1.21	($\rightarrow$ 0, $\rightarrow$ 0)	1.00

	$\phi_1 = 0.9$		$\phi_1 = 0.5$		$\phi_1 = 0.1$	
θ_1	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2	$\boldsymbol{\alpha}_{opt}^T = (\alpha_1, \alpha_2)$	MSE/σ^2
0.90	(1.00, $\rightarrow$ 0)	1.81	(1.00, $\rightarrow$ 0)	1.81	(1.00, $\rightarrow$ 0)	1.82
0.70	(1.00, $\rightarrow$ 0)	1.49	(1.00, $\rightarrow$ 0)	1.52	(0.98, $\rightarrow$ 0)	1.56
0.50	(1.00, $\rightarrow$ 0)	1.26	(1.00, $\rightarrow$ 0)	1.33	(0.17, $\rightarrow$ 0)	1.36
0.30	(1.00, $\rightarrow$ 0)	1.12	(0.98, $\rightarrow$ 0)	1.25	($\rightarrow$ 0, $\rightarrow$ 0)	1.16
0.10	(1.00, $\rightarrow$ 0)	1.05	(0.67, $\rightarrow$ 0)	1.21	($\rightarrow$ 0, $\rightarrow$ 0)	1.04

- N_t is an ARIMA$(0,1,q)$ with $0 \leq q \leq 2$. The optimal value of α_1 is equal to 1 when $\theta_1 > 0$ and less than 1 when $\theta_1 < 0$, and the optimal value of α_2 is often extremely small ($\rightarrow 0$). Holt's method performs well when $\theta_1 < 0$ and $|\theta_2|$ is small.

- N_t is an ARIMA$(0,0,q)$ with $0 \leq q \leq 2$. The optimal value of both α_1 and α_2 are often extremely small ($\rightarrow (0,0)^T$). Holt's method performs well both $|\theta_1|$ and $|\theta_2|$ are small.

- N_t is an ARIMA$(1,2,0)$. The optimal value of both α_1 and α_2 are large, and overestimation of either α_1 or α_2 is less serious than the equivalent underestimation. Holt's method performs well when $|\phi_1|$ is small, say $-0.8 < \phi_1 < 0.4$.

- N_t is an ARIMA$(1,1,0)$. The optimal value of α_1 equals to 1 when $0 < \phi_1 < 1$ and increases with ϕ_1 when $-1 < \phi_1 < 0$. The optimal value of α_2 is extremely small ($\rightarrow 0$) when $-1 < \theta_1 \leq 1/3$ and otherwise increases as ϕ_1 increases. Holt's method performs well except when ϕ_1 is close to -1.

- N_t is an ARIMA$(1,0,0)$. The optimal value of α_1 is extremely small ($\rightarrow 0$) when $-1 < \theta_1 \leq 1/3$ and otherwise increases as ϕ_1 increases. The optimal value of α_2 stays at an extremely small value ($\rightarrow 0$). Holt's method performs well except when ϕ_1 is close to -1.

- N_t is an ARIMA$(1,2,1)$. The optimal value of both α_1 and α_2 are large. Holt's method performs well when either $-1 < \theta_1 < 0$ and $|\phi_1|$ is small or $\theta_1\phi_1 < 0$ and $|\theta_1| \approx |\phi_1|$.

- N_t is an ARIMA$(1,1,1)$. The optimal value of α_1 is often large while the optimal value of α_2 is often very small. Holt's method performs well when either $-1 < \theta_1 < 0$ and $|\phi_1|$ is small or $\theta_1\phi_1 < 0$ and $|\theta_1| \approx |\phi_1|$

- N_t is an ARIMA$(1,0,1)$. The optimal values of both α_1 and α_2 are often small. Holt's method performs well when $\theta_1\phi_1 < 0$ and $|\theta_1| \approx |\phi_1|$

As a result, N_t is an ARIMA$(p,2,q)$ with $0 \leq p \leq 1$ and $0 \leq q \leq 2$, the optimal values of both α_1 and α_2 tend to be large, and overestimation of either α_1 or α_2 is less serious than the equivalent underestimation. In addition, Holt's method performs well for an IMA$(2,q)$ with $\theta_1 < 0$ and $0 < \theta_2 < 1$, an ARI$(1,2)$ with small $|\phi_1|$, and an ARIMA$(1,2,1)$ with $-1 < \theta_1 < 0$ and $|\phi_1|$ small. When N_t is an ARIMA$(p,1,q)$ with $0 \leq p \leq 1$ and $0 \leq q \leq 2$, the optimal value of α_1 tends to be large while the optimal value of α_2 tends to be small, and often overestimation of α_1 is less serious than the equivalent underestimation. In addition, Holt's method performs well for an IMA$(1,q)$ with $\theta_1 < 0$ and $|\theta_2|$ small, an ARI$(1,1)$ with large ϕ_1, and an ARIMA$(1,1,1)$ with $-1 < \theta_1 < 0$ and $|\phi_1|$ small. When N_t is an ARMA(p,q) with $0 \leq p \leq 1$ and $0 \leq q \leq 2$, the optimal values of both α_1 and α_2 are often extremely small. Holt's method performs well for an MA(q) with small $|\theta_1|$ and $|\theta_2|$, an AR(1) with large ϕ_1 is large, and an ARMA$(1,1)$ with $\theta_1\phi_1 < 0$ and $|\theta_1| \approx |\phi_1|$.

CHAPTER IV

EXPONENTIAL SMOOTHING WITH COVARIATES

4.1 Introduction

All of the ES methods described in Chapter 2), regardless of how they model a time series, an additive trend with a multiplicative seasonality or a damped multiplicative trend without seasonalities or any other forms (see Table 2.1 in Chapter 2), implicitly assume that past observations of a time series contain all information required for forecasting its future. In other words, all of these methods forecast the future of a time series using only its past observations. The history of a time series certainly contains knowledge about its future. While, other information beyond what is available in a series' own history may also shed light on the series' movements along time and therefore lead to more accurate forecasting of its future if incorporated. For example, Figure 4.1(a) displays the number of people (in thousands) killed due to motor vehicle accidents in the United States from 1911 to 1970. Supposing that we are at the beginning of the year 1971 and would like to forecast the number of motor vehicle deaths in the next three years, we could decompose the death series into two components, a trend and a seasonality,

$$deaths = trend + seasonality + disturbance\ term \tag{4.1}$$

and choose an appropriate ES method for forecasting. On the other hand, the number of motor vehicle deaths may be affected by various factors such as annual vehicle miles of travel, quality of roads, behaviors of drivers, and conditions of weather. Figure 4.1(b) shows annual vehicle miles of travel (in billions) in the United States from 1911

to 1970. The mile series exhibits movements that appear correlated to those of the death series. For instance, the two series both dropped suddenly in 1942 and then started to climb rapidly in 1944. This suggests a regression model

$$deaths = miles + road + driver + weather + disturbance\ term. \tag{4.2}$$

Using only the series' own history for forecasting, model (4.1) might lose valuable information contained in the influencing factors. Model (4.2), on the other hand, consider only influencing factors that may not be able to completely explain the movements of the death series.Furthermore, although some factors such as miles of travel are easy to measure, some factors such as the quality of roads and the behaviors of drivers are difficult to quantify. A possible solution is to take advantage of all available information, using ES methods to model what are left unexplained by measurable factors in the movements of the time series being forecasted.

From now on, we will refer to those measurable factors as "covariates."

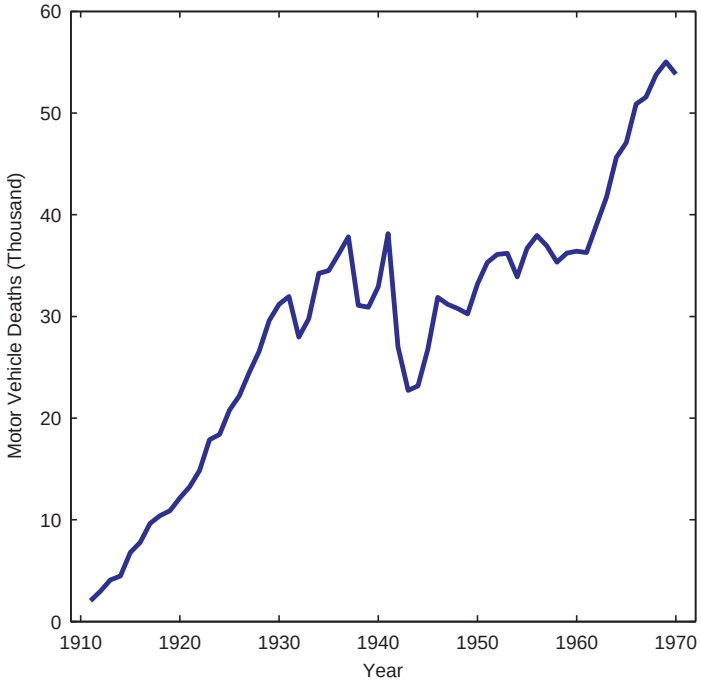
4.2 Exponential Smoothing with Covariates (ESCov)

Let $\mathbf{z}_t, t = 1, 2, \cdots$, denote the observed values of a $q \times 1$ vector of covariates. Under the assumption of a linear relationship between Y_t and $\mathbf{z}_t$, the series of interested can be adequately represented by a model of the form

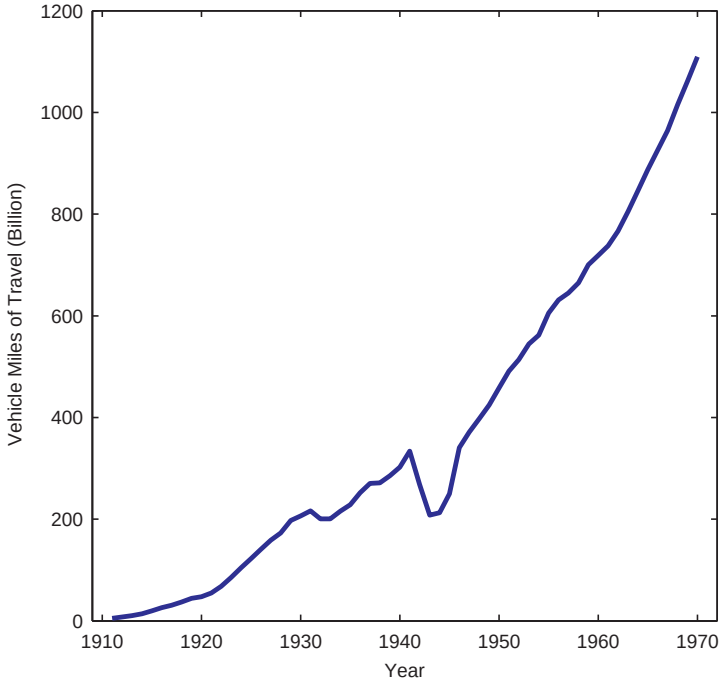
$$Y_t = \mu_t + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t, \tag{4.3}$$

where μ_t is the intercept at time t, $\boldsymbol{\delta}$ is a vector of q constant coefficients, and ϵ_t is a white noise process with $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = \sigma^2$. Model (4.3) is different from a classical linear regression model in that the intercept μ_t here is time-varying (as indicated by the index t) and more importantly is serially correlated.

A new forecasting method, ESCov, is proposed for forecasting a time series like Y_t in model (4.3). This new method estimates and forecasts the time-varying and serially



(a)



(b)

Figure 4.1: US Motor Vehicle Deaths and Miles of Travel from 1911 to 1970

correlated intercept μ_t using ES methods, thus the name "Exponential Smoothing with Covariates." ESCov has two advantages by taking into account both the history of the time series of interest and the information hidden in quantifiable covariates. First, incorporating covariates may lead to a more accurate forecast of the time series of interest compared to considering only the series' past values. Second, to forecast, we only need to know how the time series of interest moves along time. However, to make the series move in the desired directions, for example to reduce deaths due to motor vehicle accidents, we need to understand the reasons behind its movements. Such knowledge can often be learned from its relationship with other variables.

4.2.1 The Procedure

Let $l_{t+h|t}$ denote the h-step-ahead forecast of μ_{t+h} made at time t. Let $\mathbf{d}$ denote the estimator of $\boldsymbol{\delta}$. The h-step-ahead forecast of Y_{t+h} at time t by ESCov is

$$\hat{Y}_{t+h|t} = l_{t+h|t} + \mathbf{d}^T \mathbf{z}_{t+h}, \quad h > 0, \tag{4.4}$$

and the corresponding h-step-ahead forecast error is

$$e_{t+h|t} = Y_{t+h} - \hat{Y}_{t+h|t} = Y_{t+h} - l_{t+h|t} - \mathbf{d}^T \mathbf{z}_{t+h}. \tag{4.5}$$

The h-step-ahead forecast of the intercept, $l_{t+h|t}$, depends on which ES method is used to estimate and forecast the intercept μ_t. Suppose that Holt's method is chosen, then

$$l_{t+h|t} = l_t + hb_t, \tag{4.6}$$

in which l_t and b_t are calculated using the recurrence equations

$$l_t = l_{t-1} + b_{t-1} + \alpha_1 e_t, \tag{4.7a}$$

$$b_t = b_{t-1} + \alpha_2 e_t, \tag{4.7b}$$

where α_1 and α_2 are smoothing parameters taking values in the interval (0,1], and e_t is the one-step-ahead forecast error from ESCov

$$e_t \equiv e_{t|t-1} = Y_t - l_{t-1} - b_{t-1} - \mathbf{d}^T \mathbf{z}_t. \tag{4.8}$$

When the future values of covariates, $\mathbf{z}_{t+h}$, are unknown, for instance the number of miles driven in future in the motor vehicle deaths example is unknown, their predictions $\hat{\mathbf{z}}_{t+h|t}$ are used instead.

4.2.2 A General Form

Any ES method could be chosen for estimating and forecasting the intercept μ_t. A general form of ESCov consists of two equations,

- the recurrence equation

$$\mathbf{b}_t = g(\mathbf{b}_{t-1}) + w(\boldsymbol{\alpha}, \mathbf{b}_{t-1})e_t \tag{4.9}$$

- and the forecasting equation

$$\hat{Y}_{t+h|t} = f(\mathbf{b}_t) + \mathbf{d}^T\mathbf{z}_{t+h}, \quad h > 0, \tag{4.10}$$

where $\mathbf{b}_t$ is a $p \times 1$ vector of smoothed statistics (e.g., l_t and b_t in Holt's method); $\boldsymbol{\alpha}$ is a $p \times 1$ vector of smoothing parameters with $0 < \alpha_i \leq 1, i = 1, \cdots, p$; g and w are mappings from $\Re^p$ to $\Re^p$; and f is a mapping from $\Re^p$ to $\Re$.

4.2.3 Parameters Estimation

To use ESCov for forecasting, we have to find appropriate values for $\mathbf{b}_0$ and parameters $\boldsymbol{\alpha}$ and $\mathbf{d}$ (and the damping factor ϕ if damped ES methods are chosen). Then, the value of $\mathbf{b}_t$, $t > 0$, can be obtained recursively using the recurrence equation (4.9), and forecasts for future values can be calculated using the forecasting equation (4.10).

• Estimation of $\mathbf{b}_0$

A heuristic procedure is used to estimate $\mathbf{b}_0$. Let T_0 be a positive integer, and, unless particularly mentioned, T_0 is set to ten.

a). For non-seasonal data, fit a linear regression model with a deterministic time trend $a(t)$ and the predictor vector $\mathbf{z}_t$ on the first $k \in \{p+q,\ p+q+1,\ \cdots,\ T_0\}$

observations and compute the ordinary least squares (OLS) estimates of the regression coefficients (including the intercept). Then, set $\mathbf{b}_0$ to the average of the $T-p-q+1$ estimates of the coefficients of the time trend.

The deterministic time trend used in the regression model depends on the ES method chosen for the estimation of μ_t. For methods with no trend like N-N, a constant trend is included (i.e., $a(t)=a_0$), and $\mathbf{b}_0$ is set to the average of the estimates of a_0. For methods with additive trend like A-N and DA-N, a linear trend is used (i.e., $a(t)=a_0+a_1t$), and $\mathbf{b}_0$ is set to the average of the estimates of $(a_0,a_1)^T$.

b). For seasonal data, first estimate the initial seasonal indices, $c_0,c_1,\cdots,c_{-M+1}$, then apply the procedure above to the deseasonalized data. The estimated time trend coefficients together with the estimated seasonal indices form $\mathbf{b}_0$.

To estimate the initial seasonal indices, compute a M-moving average using the first few seasonal cycles, de-trend the data by dividing Y_t by (for multiplicative seasonality) or subtracting from Y_t (for additive seasonality) the moving averages, and average the de-trended data for each season to give the initial seasonal indices, $c_0,c_1,\cdots,c_{-M+1}$. If data are available, four complete seasonal cycles are used for the estimation of initial seasonal indices.

• Estimation of Parameters $\boldsymbol{\alpha}$ and $\mathbf{d}$

The parameters $\boldsymbol{\alpha}$ and $\mathbf{d}$ (and ϕ) are estimated by minimizing the sum of squared one-step-ahead forecast errors

$$SSE=\sum_{t=1}^{T}(Y_t-Y_{t|t-1})^2, \tag{4.11}$$

where T is the number of observations in the training data.

4.3 Numerical Experiments

Before beginning with numerical experiments, we introduce four accuracy measures that will be used to compare the performance of different forecasting methods.

4.3.1 Four Accuracy Measures

Let Y_t denote the actual observation at time t, and $\hat{Y}_t$ the forecast. The four accuracy measures used are

- Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100 \tag{4.12}$$

- symmetric Mean Absolute Percentage Error (sMAPE)

$$\text{sMAPE} = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{Y_t - \hat{Y}_t}{(Y_t + \hat{Y}_t)/2}\right| \times 100 \tag{4.13}$$

- Median Absolute Percentage Error (MdAPE)

$$\text{MdAPE} = \text{median}\{\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|,\ t = 1, 2\cdots, T\} \times 100 \tag{4.14}$$

- Root Mean Square Error (RMSE)

$$\text{RMSE} = \left[\frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2\right]^{1/2} \tag{4.15}$$

Among them, MAPE, sMAPE, and MdAPE are unit free while RMSE has the same unit as Y_t.

4.3.2 Two Examples

Example 1. UK Annual Consumption of Spirits. The data set contains 69 observations on the logarithms of three variables, consumption of spirits per capita (Y_t), real income per capita ($z_{1,t}$), and relative price of spirits ($z_{2,t}$) (deflated by cost-of-living index), from 1870 to 1938 in the United Kingdom. Figure 4.2 shows plots

of these three series. This data have frequently appeared in literature since first studied by Prest (1949) and are often analyzed by fitting a linear regression model with a deterministic linear or quadratic time trend and a first-order autoregressive error (Fuller 1996).

The consumptions of spirits were forecasted using the following five methods:

1). ESCov1 – ESCov having two covariates, income and price

2). ESCov2 – ESCov having one covariate, income

3). ESCov3 – ESCov having one covariate, price

4). Holt's method

5). A linear regression model with a deterministic linear time trend and an AR(1) error

$$Y_t = \beta_0 + \beta_1 z_{1,t} + \beta_2 z_{2,t} + \beta_3 z_{3,t} + U_t, \tag{4.16a}$$

$$U_t = \varphi U_{t-1} + \epsilon_t, \tag{4.16b}$$

where $z_{3,t} = (t - 1869)/100$, where t denote the year.

For ESCov1, ESCov2, and ESCov3, Holt's method was chosen to estimate the intercept μ_t. The regression model with an AR(1) error was fitted using the SAS procedure AUTOREG.

The spirit data set was divided into two portions. The 59 observations from 1870 to 1928 formed the training data for parameter estimation, and forecasts for the following 10 years, 1929–1938, were generated. Figure 4.3 shows forecasts by those five methods. All of them performed well for short-term forecasts, say $h < 4$. For long-term forecasts, ESCov's did better than Holt's method and the regression model did. Holt's method tended to under-predict spirit consumptions while the regression model tended to overshot the data.

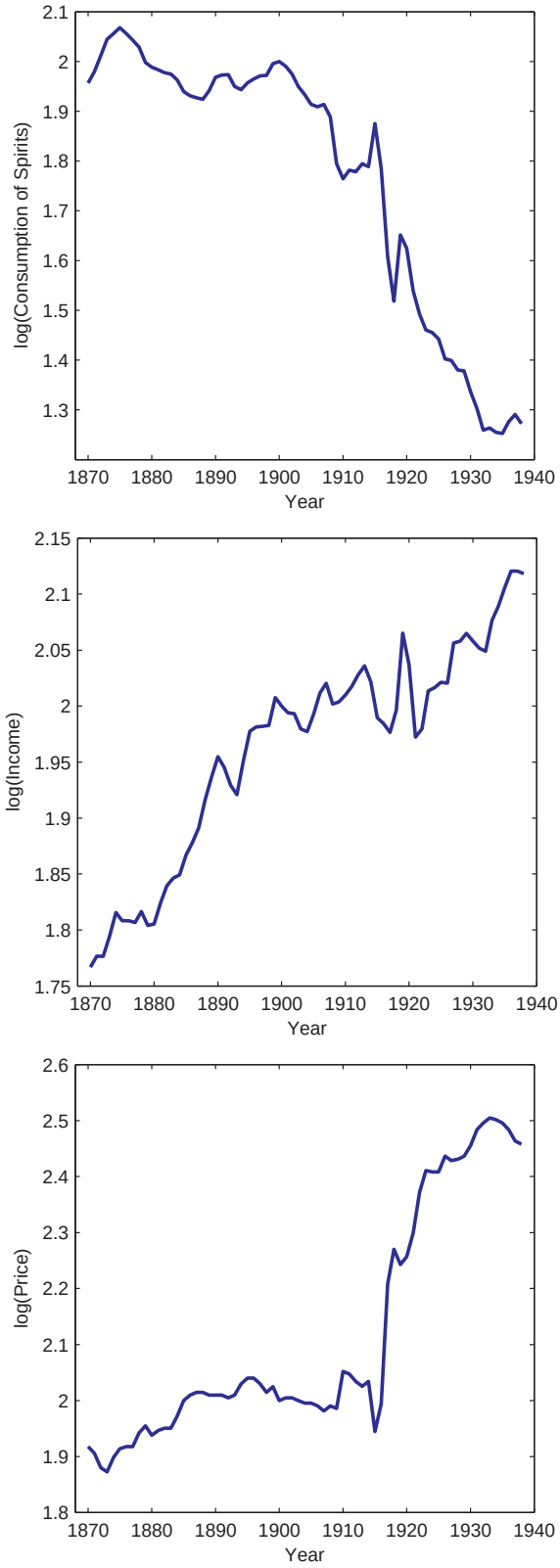


Figure 4.2: UK Spirit Consumptions per Capita, Income per Capita and Price of Spirits from 1870 to 1938.

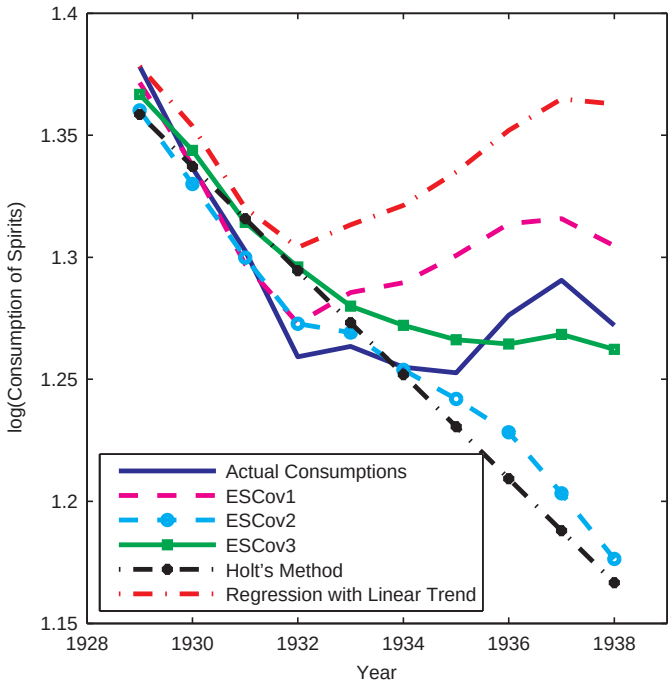


Figure 4.3: UK Consumptions of Spirits per Capita and Forecasts for 1929–1938

For a better comparison among ESCov's and Holt's method (notice that ESCov reduces to Holt's method when there are no covariates and the intercept μ_t is estimated and forecasted using Holt's method), the following procedure was employed. First, the 50 observations from 1870 to 1919 were used as the training data to estimate the parameters, and the forecasts of spirit consumptions were generated for the following ten years from 1920 to 1929. Then, the next observation (i.e., the observation in 1920) was added to the training data, the parameters were re-estimated, and new ten-year forecasts were generated. Repeating this procedure until all but the last ten observations were included in the training data gave 10 h-step-ahead forecasts for $h = 1, 2, \cdots, 10$. Last, the 10 forecasts for each h were used to calculate MAPE, sMAPE, MdAPE and RMSE.

Table 4.1 shows that ESCov2 and ESCov3 produced more accurate forecasts than ESCov1 and Holt's method did, having smaller MAPE, sMAPE and RMSE for all 10 horizons (i.e., $h = 1, 2, \cdots, 10$) and smaller MdAPE for 8 horizons (i.e., $h =$

$1, 2, 3, 4, 5, 7, 8, 10$). For $h = 6$ and $h = 9$, Holt's method gave the smallest MdAPE as shown by numbers in bold face. In general, ESCov1 had the poorest performance in terms of MAPE, sMAPE, MdAPE and RMSE, which implies that including more covariates do not necessarily leads to more accurate forecasts. In addition, it can be seen from Figure 4.3 and Table 4.1 that forecast accuracy, as expected, decreases as the forecast horizon h increases.

Example 2. US Annual Motor Vehicle Deaths. The data set includes 60 observations from 1911 to 1970 on the number of deaths (in thousand) due to motor vehicle accidents and the number of miles (in billion) driven in the United States. As mentioned at the beginning of this chapter, many factors may contribute to motor vehicle deaths. However, not all of them are available or quantifiable. In this example, only annual miles driven is included in ESCov as the covariate.

Using the first 45 observations from 1911 to 1955 as the training data, we generated forecasts for the last 15 years using the follow methods

1). ESCov. Holt's method was chosen to estimate the intercept μ_t, and two sets of forecasts of motor vehicle deaths from 1956 to 1970 were obtained. One used the real values of annual miles driven from 1956 to 1970 while the other used the predicted annual miles driven for 1956–1970 by damped Holt's method.

2). Holt's method

3). Regression with an AR(1) error. Two linear regression models with an AR(1) error

$$U_t = \varphi U_{t-1} + \epsilon_t \tag{4.17}$$

were fitted. One assumed a linear time trend

$$Y_t = \beta_0 + \beta_1 z_t + \beta_2 t + U_t \tag{4.18}$$

Table 4.1: UK per capita Consumption of Spirits Forecasts for 1925-1938

MAPE										
Horizon (h)	1	2	3	4	5	6	7	8	9	10
ESCov1	1.3366	2.3454	2.8955	3.3398	4.0707	4.9602	5.9875	6.5014	7.1343	7.7698
ESCov2	1.1786	1.6593	1.9935	2.1152	2.2478	2.7187	3.0760	3.6116	4.1992	5.1664
ESCov3	0.8001	1.2755	1.5371	2.0394	2.3576	2.8622	3.3288	3.7000	3.9710	4.0496
Holt's Method	1.4081	2.1857	2.5879	2.9458	3.1254	2.9621	3.5663	3.8150	4.3982	5.6591
sMAPE										
Horizon (h)	1	2	3	4	5	6	7	8	9	10
ESCov1	0.3334	0.5851	0.7186	0.8207	0.9946	1.2047	1.4494	1.5644	1.7085	1.8515
ESCov2	0.2935	0.4105	0.4905	0.5189	0.5507	0.6650	0.7498	0.8841	1.0361	1.2808
ESCov3	0.1984	0.3149	0.3797	0.5031	0.5800	0.7019	0.8133	0.9014	0.9665	0.9821
Holt's Method	0.3495	0.5374	0.6327	0.7196	0.7662	0.7234	0.8712	0.9377	1.0929	1.4189
MdAPE										
Horizon (h)	1	2	3	4	5	6	7	8	9	10
ESCov1	0.7084	2.5208	3.0522	4.0524	4.3166	5.4960	6.5742	7.4992	9.1791	9.6750
ESCov2	1.1547	1.5742	1.1114	0.9450	1.2239	2.1482	1.7648	2.8174	3.9334	5.5346
ESCov3	0.6264	0.7651	1.2669	1.5396	1.9491	2.6952	2.7439	3.2357	3.7863	4.3639
Holt's Method	1.3229	1.6874	1.2853	1.8597	2.4089	**1.1431**	2.2659	2.8573	**3.4226**	6.1256
RMSE										
Horizon (h)	1	2	3	4	5	6	7	8	9	10
ESCov1	0.0250	0.0412	0.0488	0.0531	0.0640	0.0753	0.0849	0.0931	0.1020	0.1112
ESCov2	0.0199	0.0324	0.0390	0.0418	0.0439	0.0512	0.0567	0.0602	0.0669	0.0777
ESCov3	0.0190	0.0279	0.0287	0.0333	0.0384	0.0458	0.0533	0.0593	0.0628	0.0664
Holt's Method	0.0275	0.0465	0.0569	0.0598	0.0593	0.0640	0.0688	0.0691	0.0757	0.0889

while the other assumed a quadratic time trend

$$Y_t = \beta_0 + \beta_1 z_t + \beta_2 t + \beta_3 (t - 30)^2 + U_t, \tag{4.19}$$

where z_t denotes the annual miles driven, and 1910 is the origin for t. The regression models were fitted using the SAS procedure AUTOREG, and the real values of annual miles driven from 1956 to 1970 were used for forecasting.

Figure 4.4 shows that the forecasts from ESCov's were closer to the actual deaths than these from Holt's method and the two regression models. In average, ESCov with real miles outperformed all the other methods. The poor performance of ESCov with predicted miles for large forecast horizon was due to the under-predictions of miles by damped Holt's method (see Figure 4.5). The two regression models performed worst. Although the regression model with a quadratic time trend performed well for small forecast horizons, say $h < 5$, its performance worsened quickly as the forecast horizon increased.

Following the same procedure as described in Example 1, we obtained 10 forecasts for each of 10 horizons using ESCov and Holt's method. Table 4.2 shows that, ESCov with real miles had the smallest MAPE, sMAPE and RMSE except when $h = 3$, for which ESCov with predicted miles gave smaller MAPE, sMAPE and RMSE. ESCov with predicted miles outperformed Holt's method except when the forecast horizon was large, say $h > 7$. This was as mentioned before due to the under-predictions of miles by damped Holt's method for large horizons.

4.4 Statistical Properties

In this section, we investigate the statistical properties of ESCov. We identify SSOE state space models underlying ESCov, discuss the maximum likelihood estimation of unknown parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, propose a model section procedure for ESCov, and derive the variances of h-step-ahead forecasts by ESCov.

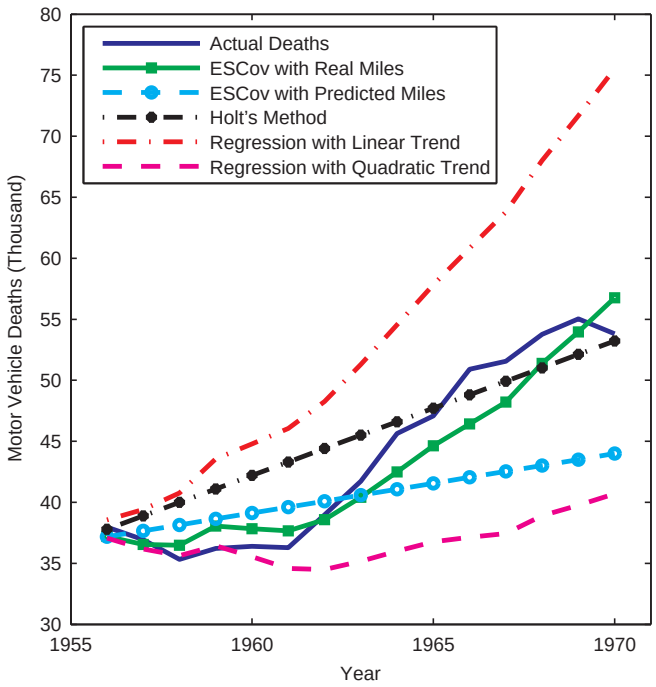


Figure 4.4: US Motor Vehicle Deaths and Forecasts for 1956–1970

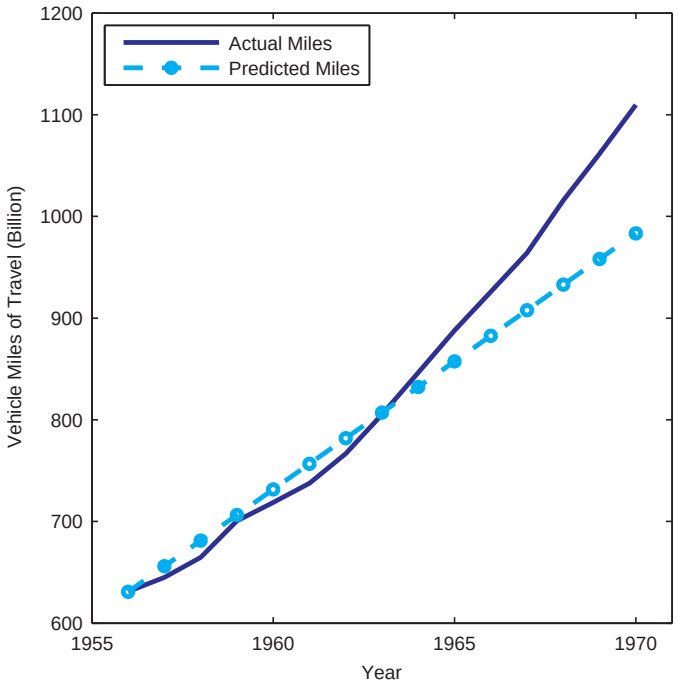


Figure 4.5: US Vehicle Miles of Travel and Forecasts for 1956–1970

Table 4.2: US Motor Vehicle Death Forecasts for 1951-1970

MAPE

Horizon (h)	1	2	3	4	5	6	7	8	9	10
ESCov with Real Miles	2.4227	4.7562	6.3842	7.4074	8.1681	9.3893	10.1389	11.0981	10.9241	9.5654
ESCov with Predicted Miles	3.3353	4.7651	5.9768	7.4242	9.9475	11.7139	12.1721	14.3293	16.3201	17.8039
Holt's Method	3.8679	6.3261	7.6475	9.2213	11.3274	13.4502	14.0270	13.4996	12.8357	11.5241

sMAPE

Horizon (h)	1	2	3	4	5	6	7	8	9	10
ESCov with Real Miles	0.5962	1.1640	1.5678	1.8439	2.0335	2.3262	2.4889	2.7140	2.6685	2.3338
ESCov with Predicted Miles	0.8300	1.1895	1.5039	1.9168	2.6090	3.1578	3.3724	4.0267	4.6544	5.1365
Holt's Method	0.9434	1.5221	1.8219	2.2119	2.7026	3.2220	3.3608	3.2547	3.1115	2.8071

MdAPE

Horizon (h)	1	2	3	4	5	6	7	8	9	10
ESCov with Real Miles	1.6667	4.7636	6.6080	6.9327	9.1203	12.3475	11.1495	11.9346	8.5822	8.7934
ESCov with Predicted Miles	2.2624	3.2958	5.8991	6.0159	7.7923	7.9660	7.5885	11.5464	15.6337	17.7483
Holt's Method	3.0543	5.1177	6.9275	9.1603	9.9504	13.9313	14.1878	12.9546	13.2113	12.5358

RMSE

Horizon (h)	1	2	3	4	5	6	7	8	9	10
ESCov with Real Miles	1.0967	2.0165	2.6554	3.2032	3.7492	4.6116	5.1975	5.4751	5.3339	4.8728
ESCov with Predicted Miles	1.5487	2.2438	2.6301	3.7923	5.1597	6.6409	7.6883	9.0695	10.3543	11.3020
Holt's Method	1.7269	2.6322	3.2238	3.8475	4.6996	5.6325	6.0633	6.3517	6.4796	6.0960

4.4.1 Underlying Statistical Models

As discussed in Chapter 2, ES methods have SSOE state space models as their underlying statistical models. Similarly, an SSOE state space model underlying ESCov can also be identified. Such an SSOE state space model consists of an observation equation

$$Y_t = f(\boldsymbol{\beta}_{t-1}) + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t \tag{4.20}$$

and a transition equation

$$\boldsymbol{\beta}_t = g(\boldsymbol{\beta}_{t-1}) + w(\boldsymbol{\alpha}, \boldsymbol{\beta}_{t-1})\epsilon_t, \tag{4.21}$$

where $\boldsymbol{\beta}_t$ is a $p \times 1$ state vector, denoting the true value of $\mathbf{b}_t$ in equations (4.9) and (4.10).

According to the observation equation (4.20), the one-step-ahead forecast made at time $t-1$, given that $\boldsymbol{\beta}_{t-1}$, $\boldsymbol{\delta}$, and $\mathbf{z}_t$ are known, is

$$\hat{Y}_{t|t-1} = f(\boldsymbol{\beta}_{t-1}) + \boldsymbol{\delta}^T \mathbf{z}_t, \tag{4.22}$$

and the corresponding one-step-ahead forecast error is

$$e_t = Y_t - \hat{Y}_{t|t-1} = \epsilon_t. \tag{4.23}$$

As a result, the transition equation (4.21) is the same as the recurrence equation (4.9) except that the vector of smoothed statistics, $\mathbf{b}_t$, in (4.9) is replaced by its true value, $\boldsymbol{\beta}_t$, in (4.21). In another words, given the true initial value $\boldsymbol{\beta}_0$ and the true values of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ (and ϕ), ESCov will produce optimal (i.e., minimum mean squared error) forecasts for a time series generated from the model in (4.20) and (4.21).

Table 4.3 lists SSOE state space models that underpin ESCov with different ES methods chosen to estimate and forecast the intercept μ_t. All those SSOE models are homoscedastic if ϵ_t in the observation equation is assumed to be a white noise process

with zero mean (i.e., $E[\epsilon_t] = 0$) and constant variance (i.e., $E[\epsilon_t^2] = \sigma^2$). Replacing ϵ_t by $u(\boldsymbol{\beta}_{t-1}, \mathbf{z}_t)\epsilon_t$, where u is a mapping from $\Re^{p+q}$ to $\Re$, and assuming that

$$E[\epsilon_t \boldsymbol{\beta}_k] = 0, \text{ for } k < t \tag{4.24}$$

and

$$E[\epsilon_t \mathbf{z}_k] = 0, \text{ for all } k \tag{4.25}$$

give heteroscedastic underlying SSOE state space models for ESCov.

4.4.2 Maximum Likelihood Estimation

Knowing underlying statistical models for ESCov makes the maximum likelihood estimation of unknown parameters possible. Given the data $\{Y_t, \mathbf{z}_t; t = 1, 2, \cdots, T\}$, we hope to obtain the maximum likelihood estimators of unknown parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$. The SSOE model in (4.20) and (4.21) implies that the behavior of the time series of interest depends on not only parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ but also the initial state $\boldsymbol{\beta}_o$. Ord et al. (1997) proposed a conditional maximum likelihood approach that bases the parameter estimation upon the likelihood conditional on the initial state $\boldsymbol{\beta}_0$.

Let $p(\cdot)$ denote a probability function. The likelihood function conditional on $\boldsymbol{\beta}_0$, given the data $\{Y_t, \mathbf{z}_t; t = 1, 2, \cdots, T\}$, is

$$L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \sigma^2 | \boldsymbol{\beta}_0) = p(Y_1, \cdots, Y_T; \boldsymbol{\alpha}, \boldsymbol{\delta}, \sigma^2 | \boldsymbol{\beta}_0). \tag{4.26}$$

The probability law, $P(AB) = P(A|B)P(B)$, implies that

$$L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \sigma^2 | \boldsymbol{\beta}_0) = \prod_{t=1}^{T} p(Y_t; \boldsymbol{\alpha}, \boldsymbol{\delta}, \sigma^2 | Y_1, \cdots, Y_{t-1}, \boldsymbol{\beta}_0). \tag{4.27}$$

Under the assumption that the underlying SSOE state space model is homoscedastic (i.e., $u(\boldsymbol{\beta}_{t-1}, \mathbf{z}_t) = 1$) and the white noise process ϵ_t is Gaussian, the conditional likelihood function becomes

$$L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \sigma^2 | \boldsymbol{\beta}_0) = (2\pi\sigma^2)^{-T/2} \cdot \exp(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \epsilon_t^2). \tag{4.28}$$

Table 4.3: Underlying SSOE State Space Models for ESCov. (**N** - None, **A** - Additive, **M** - Multiplicative, **DA** - Damped Additive)

	Seasonality		
Trend	**N**	**A**	**M**
N	$Y_t = \mu_{t-1} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \alpha_1 \epsilon_t$	$Y_t = \mu_{t-1} + s_{t-M} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \alpha_1 \epsilon_t$ $s_t = s_{t-M} + \alpha_3 \epsilon_t$	$Y_t = \mu_{t-1} s_{t-M} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \alpha_1 \epsilon_t / s_{t-M}$ $s_t = s_{t-M} + \alpha_3 \epsilon_t / \mu_{t-1}$
A	$Y_t = \mu_{t-1} + \beta_{t-1} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \beta_{t-1} + \alpha_1 \epsilon_t$ $\beta_t = \beta_{t-1} + \alpha_2 \epsilon_t$	$Y_t = \mu_{t-1} + \beta_{t-1} + s_{t-M} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \beta_{t-1} + \alpha_1 \epsilon_t$ $\beta_t = \beta_{t-1} + \alpha_2 \epsilon_t$ $s_t = s_{t-M} + \alpha_3 \epsilon_t$	$Y_t = (\mu_{t-1} + \beta_{t-1}) s_{t-M} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \beta_{t-1} + \alpha_1 \epsilon_t / s_{t-M}$ $\beta_t = \beta_{t-1} + \alpha_2 \epsilon_t / s_{t-M}$ $s_t = s_{t-M} + \alpha_3 \epsilon_t / (\mu_{t-1} + \beta_{t-1})$
DA	$Y_t = \mu_{t-1} + \phi\beta_{t-1} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \phi\beta_{t-1} + \alpha_1 \epsilon_t$ $\beta_t = \phi\beta_{t-1} + \alpha_2 \epsilon_t$	$Y_t = \mu_{t-1} + \phi\beta_{t-1} + s_{t-M} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \phi\beta_{t-1} + \alpha_1 \epsilon_t$ $\beta_t = \phi\beta_{t-1} + \alpha_2 \epsilon_t$ $s_t = s_{t-M} + \alpha_3 \epsilon_t$	$Y_t = (\mu_{t-1} + \phi\beta_{t-1}) s_{t-M} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t$ $\mu_t = \mu_{t-1} + \phi\beta_{t-1} + \alpha_1 \epsilon_t / s_{t-M}$ $\beta_t = \phi\beta_{t-1} + \alpha_2 \epsilon_t / s_{t-M}$ $s_t = s_{t-M} + \alpha_3 \epsilon_t / (\mu_{t-1} + \phi\beta_{t-1})$

Therefore, the conditional maximum likelihood estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, when the white noise process is Gaussian, are equivalent to their minimum sum of squared one-step-ahead forecast error estimators (see equation (4.11)).

4.4.3 Model Selection

Model selection for ESCov addresses two questions: which covariates should be incorporated into ESCov, and which ES method should be used for the estimation of the time-varying intercept μ_t.

There are two general approaches to model selection, hypothesis testing and criterion optimization. The hypothesis tests are often the choice in econometric model building and can be set up in two opposite ways, leading to two types of tests, unit root tests and stationary tests. Unit root tests, formulated in an autoregressive framework, have the null hypothesis that a time series has a unit root while the alternative that the time series is stationary. On the contrary, stationary tests, based on models in state space form, have the null hypothesis that a component (e.g., level, trend, or seasonality) is deterministic while the alternative that it is stochastic and nonstationary. A detailed review on unit root and stationary tests was given by Harvey (2005). All of existing stationary tests are for linear MSOE state space models. The generalization to linear SSOE state space models is possible while the generalization to SSOE models for multiplicative trend or seasonality is another story.

Different from hypothesis testing, the criterion-based approach selects from potential candidate models the one that minimizes the selected criterion. Commonly used criteria include Akaike's information criterion (AIC) (Akaike 1973), a modified AIC (Hurvich and Tsai 1989), Bayesian information criterion (BIC) (Schwarz 1973), HQ (Hannan and Quinn 1979), to name a few. With so many criteria available, which one to use is not a surprising question that is difficult to answer as no criterion is

universally superior to all the others. In the context of exponential smoothing, a comparison of six different criteria based on simulated and real-life time series by Billah et al. (2006) revealed that AIC was the best for selecting from three ES methods, N-N, A-N, and A-A.

Model selection for ESCov, which has underlying models in SSOE state space form, could use either approach. As existing hypothesis tests are for only linear MSOE state space models, the criterion-based approach is preferred since it works for any types of models as long as the selected criterion are calculable. We propose a model selection procedure that consists of two steps: preliminary analysis and criterion-based auto selection. In the preliminary analysis, domain knowledge and time series plots are used to select potential covariates and ES methods, based on which possible candidate models are formulated and used as the input to the criterion-based auto selection, which chooses from those candidates the one that minimize the selected criterion.

• Preliminary analysis

Preliminary analysis identifies potential covariates and possible ES methods based on domain knowledge and time series plots. The importance of domain knowledge in covariate selection is quite obvious. For example, the price of a product undoubtedly has an influence on its sale and should be included in ESCov as a covariate. Time series plots also plays an important role in model selection. Comparing the plot of a covariate with that of the dependent variable may indicate which of the movements in the dependent variable are capable of being explained by the covaraite. For example, Figure 4.1 shows that the death series and the mile series exhibit similar movements. Both series increased globally from 1910 to 1970 but experienced a temporary, abrupt drop in 1940. In addition, time series plots can also give hints which ES method to choose. For example, Figure 4.1(a) implies that the death series has a trend but no seasonality, which suggests the use of Holt's method or damped Holt's method.

• **Criterion-Based Auto selection**

The potential covariates and possible ES methods identified in the preliminary analysis may lead to several candidate models. The next step is to select from those candidates by optimizing certain criterion. We uses Akaike's information criterion (AIC)

$$\text{AIC} = -2\log L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \sigma^2|\boldsymbol{\beta}_0) + 2K, \tag{4.29}$$

where $\log L(\boldsymbol{\alpha}, \boldsymbol{\delta}, \sigma^2|\boldsymbol{\beta}_0)$ is the logarithm of the conditional likelihood, and K is the number of model parameters. With a Gaussian white noise process, AIC becomes

$$\text{AIC} = T\log(2\pi) + T\log(\sigma^2) + \frac{1}{\sigma^2}\sum_{t=1}^{T}\epsilon_t^2 + 2K, \tag{4.30}$$

Among those candidate models, the one with minimum AIC is selected.

4.4.4 Prediction Intervals

Up to this point, predictions for future values are given as single numbers, referred to as "point forecasts,", which alone are often not adequate as they tell nothing about prediction uncertainties. To show how uncertain a prediction is, one should supplement a point forecast with an interval of the form

$$\text{point forecast } \pm\ w \cdot \sqrt{\text{varaince of the point forecast}} \tag{4.31}$$

where w depends on the distribution of the point forecast as well as a pre-specified probability that the expected future value will fall into this interval. Such an interval is referred to as "prediction interval," whose construction clearly requires the variance of the point forecast.

The variances of h-step-ahead forecasts made at time t by ESCov can be derived analytically from underlying SSOE state space models, among which we consider only four types: linear models with additive error, nonlinear models with additive error, linear models with multiplicative error, and nonlinear models with multiplicative

Table 4.4: Four Classes of SSOE State Space Models Underlying ESCov

Class	Underlying Model	Error	ES Methods for ESCov
1	Linear	Additive	N-N, A-N, DA-N, N-A, A-A, DA-A
2	Nonlinear	Additive	N-M, A-M, DA-M
3	Linear	Multiplicative	N-N, A-N, DA-N, N-A, A-A, DA-A
4	Nonlinear	Multiplicative	N-M, A-M, DA-M

error (see Table 4.4). Along with the variances, the analytic formulas for the means of h-step-ahead forecasts are also given.

• Class 1 – Linear SSOE model with additive error

Models in class 1 are of the form

$$Y_t = \mathbf{x}^T \boldsymbol{\beta}_{t-1} + \boldsymbol{\delta}^T \mathbf{z}_t + \epsilon_t, \tag{4.32a}$$

$$\boldsymbol{\beta}_t = \mathbf{G}\boldsymbol{\beta}_{t-1} + \boldsymbol{\alpha}\epsilon_t. \tag{4.32b}$$

where $\mathbf{G}$ is a $p \times p$ constant matrix, $\mathbf{x}$ is a $p \times 1$ vector, and ϵ_t is serially independent with $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = \sigma_t^2$. This class contains six SSOE models in Table 4.3. They are models for N-N, A-N, DA-N, N-A, A-A, and DA-A. For each model in class 1, the p, $\boldsymbol{\beta}_t$, $\boldsymbol{\alpha}$, $\mathbf{x}$, and $\mathbf{G}$ are given in Table 4.5.

For $h \geq 1$, the conditional mean and variance of Y_{t+h} given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ are derived in Appendix A:

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mathbf{x}^T \mathbf{G}^{h-1} \boldsymbol{\beta}_t + \boldsymbol{\delta}^T \mathbf{z}_{t+h}, \tag{4.33a}$$

$$V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = \begin{cases} \sigma_{t+1}^2 & h = 1 \\ \sum_{i=0}^{h-2} (\mathbf{x}^T \mathbf{G}^i \boldsymbol{\alpha})^2 \cdot \sigma_{t+h-i}^2 + \sigma_{t+h}^2 & h > 1 \end{cases} \tag{4.33b}$$

The specific results for each of the six models in class 1 are given in Table 4.6. Notice that the conditional mean for any model is, as one would expect, the same as the point forecast from the corresponding ESCov. For example, $E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mu_t + \boldsymbol{\delta}^T \mathbf{z}_{t+h}$ for the model for N-N while $\hat{Y}_{t+h|t} = l_t + \mathbf{d}^T \mathbf{z}_{t+h}$ for ESCov with the intercept estimated by N-N (i.e., SES).

Table 4.5: Models in Class 1. For $k \geq 0$, $\mathbf{0}_k$ is a $k \times 1$ vector of zeros, and I_k is a $k \times k$ identity matrix.

	Seasonality	
Trend	**N**	**A**
N	$p = 1,\ \mathbf{x} = 1,\ \boldsymbol{\beta}_t = \mu_t$ $\mathbf{G} = 1,\ \boldsymbol{\alpha} = \alpha_1$	$p = M+1,\ \mathbf{x} = (1, \mathbf{0}_{M-1}^T, 1)^T,\ \boldsymbol{\beta}_t = (\mu_t, s_t, \cdots, s_{t-M+1})^T$ $\mathbf{G} = \begin{bmatrix} 1 & \mathbf{0}_{M-1}^T & 0 \\ 0 & \mathbf{0}_{M-1}^T & 1 \\ \mathbf{0}_{M-1} & \mathbf{I}_{M-1} & \mathbf{0}_{M-1} \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_3 \\ \mathbf{0}_{M-1}^T \end{bmatrix}$
A	$p = 2,,\ \mathbf{x} = (1,1)^T,\ \boldsymbol{\beta}_t = (\mu_t, \beta_t)^T$ $\mathbf{G} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$	$p = M+2,\ \mathbf{x} = (1, 1, \mathbf{0}_{M-1}^T, 1)^T,\ \boldsymbol{\beta}_t = (\mu_t, \beta_t, s_t, \cdots, s_{t-M+1})^T$ $\mathbf{G} = \begin{bmatrix} 1 & 1 & \mathbf{0}_{M-1}^T & 0 \\ 0 & 1 & \mathbf{0}_{M-1}^T & 0 \\ 0 & 0 & \mathbf{0}_{M-1}^T & 1 \\ \mathbf{0}_{M-1} & \mathbf{0}_{M-1} & \mathbf{I}_{M-1} & \mathbf{0}_{M-1} \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \mathbf{0}_{M-1}^T \end{bmatrix}$
DA	$p = 2,\ \mathbf{x} = (1,\phi)^T,\ \boldsymbol{\beta}_t = (\mu_t, \beta_t)^T$ $\mathbf{G} = \begin{bmatrix} 1 & \phi \\ 0 & \phi \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$	$p = M+2,\ \mathbf{x} = (1, \phi, \mathbf{0}_{M-1}^T, 1)^T,\ \boldsymbol{\beta}_t = (\mu_t, \beta_t, s_t, \cdots, s_{t-M+1})^T$ $\mathbf{G} = \begin{bmatrix} 1 & \phi & \mathbf{0}_{M-1}^T & 0 \\ 0 & \phi & \mathbf{0}_{M-1}^T & 0 \\ 0 & 0 & \mathbf{0}_{M-1}^T & 1 \\ \mathbf{0}_{M-1} & \mathbf{0}_{M-1} & \mathbf{I}_{M-1} & \mathbf{0}_{M-1} \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \mathbf{0}_{M-1}^T \end{bmatrix}$

Table 4.6: Class 1 – Conditional Mean and Variance of Y_{t+h} Given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$

For $h \geq 1$,

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = m_h + \boldsymbol{\delta}^T \mathbf{z}_{t+h}, \quad V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = \begin{cases} \sigma^2_{t+1}, & h = 1 \\ v_h + \sigma^2_{t+h}, & h > 1 \end{cases}$$

$$h^* = h(\text{mod } M), \quad I_{i,M-1} = \begin{cases} 1, & i(\text{mod } M) = M-1 \\ 0, & \text{otherwise} \end{cases}$$

m_h

	Seasonality	
Trend	**N**	**A**
N	μ_t	$\mu_t + s_{t-M+h^*}$
A	$\mu_t + \beta_t h$	$\mu_t + h\beta_t + s_{t-M+h^*}$
DA	$\mu_t + \beta_t \sum_{i=1}^{h} \phi^i$	$\mu_t + h\beta_t + s_{t-M+h^*}$

v_h

	Seasonality	
Trend	**N**	**A**
N	$\sum_{i=0}^{h-2} \alpha_1^2 \sigma^2_{t+h-i}$	$\sum_{i=0}^{h-2} [\alpha_1 + \alpha_3 \cdot I_{i,M-1}]^2 \sigma^2_{t+h-i}$
A	$\sum_{i=0}^{h-2} [\alpha_1 + \alpha_2 \cdot (i+1)]^2 \sigma^2_{t+h-i}$	$\sum_{i=0}^{h-2} [\alpha_1 + \alpha_2 \cdot (i+1) + \alpha_3 \cdot I_{i,M-1}]^2 \sigma^2_{t+h-i}$
DA	$\sum_{i=0}^{h-2} \left[\alpha_1 + \alpha_2 \cdot \sum_{j=1}^{i+1} \phi^j\right]^2 \sigma^2_{t+h-i}$	$\sum_{i=0}^{h-2} \left[\alpha_1 + \alpha_2 \cdot \sum_{j=1}^{i+1} \phi^j + \alpha_3 \cdot I_{i,M-1}\right]^2 \sigma^2_{t+h-i}$

• Class 2 – Nonlinear SSOE model with additive error

Let $\boldsymbol{\beta}_t^*$ be the vector of all of nonseasonal components in the state vector $\boldsymbol{\beta}_t$. That is, $\boldsymbol{\beta}_t = (\boldsymbol{\beta}_t^{*T}, s_t, \cdots, s_{t-M+1})^T$. Models in class 2 are of the form

$$Y_t = \mathbf{x}^{*T}\boldsymbol{\beta}_{t-1}^* \cdot s_{t-M} + \boldsymbol{\delta}^T\mathbf{z}_t + \epsilon_t, \tag{4.34a}$$

$$\boldsymbol{\beta}_t^* = \mathbf{G}^*\boldsymbol{\beta}_{t-1}^* + \boldsymbol{\alpha}^*\epsilon_t/s_{t-M}, \tag{4.34b}$$

$$s_t = s_{t-M} + \alpha_3\epsilon_t/(\mathbf{x}^{*T}\boldsymbol{\beta}_{t-1}^*). \tag{4.34c}$$

Three models in Table 4.3 belong to class 2. They are models for N-M, A-M, and DA-M. Furthermore, $\mathbf{x}^*$, $\boldsymbol{\beta}_t^*$, $\mathbf{G}^*$, and $\boldsymbol{\alpha}^*$ in model (4.34) for N-M, A-M, and DA-M are the same as $\mathbf{x}$, $\boldsymbol{\beta}_t$, $\mathbf{G}$, and $\boldsymbol{\alpha}$ in model (4.32) for N-N, A-N, and DA-N respectively.

Derived in Appendix B, the conditional mean and variance of Y_{t+h} given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ for $1 \le h \le M$ are

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mathbf{x}^{*T}(\mathbf{G}^*)^{h-1}\boldsymbol{\beta}_t^* \cdot s_{t+h-M} + \boldsymbol{\delta}^T\mathbf{z}_{t+h}, \tag{4.35a}$$

$$V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = \begin{cases} \sigma_{t+1}^2, & h = 1 \\ \sum_{i=0}^{h-2}(\mathbf{x}^{*T}(\mathbf{G}^*)^i\boldsymbol{\alpha}^*)^2 \cdot \sigma_{t+h-i}^2 \cdot s_{t+h-M}^2/s_{t+h-M-i}^2 + \sigma_{t+h}^2, & h > 1 \end{cases} \tag{4.35b}$$

The specific results for the three models in class 2 are given in Table 4.7.

When $h > M$, the derivations of the conditional mean and variance of Y_{t+h} give $\boldsymbol{\beta}_t$ involve deriving the mean and variance of the ratio of two random variables. They often do not exist. Therefore, only $1 \le h \le M$ are considered here.

• Class 3 – Linear SSOE model with multiplicative error

Replacing ϵ_t in model (4.32) by $u(\boldsymbol{\beta}_{t-1})\epsilon_t$ gives a model of the form

$$Y_t = \mathbf{x}^T\boldsymbol{\beta}_{t-1} + \boldsymbol{\delta}^T\mathbf{z}_t + u(\boldsymbol{\beta}_{t-1})\epsilon_t, \tag{4.36a}$$

$$\boldsymbol{\beta}_t = \mathbf{G}\boldsymbol{\beta}_{t-1} + \boldsymbol{\alpha} \cdot u(\boldsymbol{\beta}_{t-1})\epsilon_t. \tag{4.36b}$$

where $u(\boldsymbol{\beta}_t)$ is a mapping linear in $\boldsymbol{\beta}_t$. As a result, the SSOE model is linear. In other words, an SSOE model is said to be linear as long as it is linear in $\boldsymbol{\beta}_t$ and $\boldsymbol{\delta}$.

Table 4.7: Class 2 – Conditional Mean and Variance of Y_{t+h} Given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$

For $1 \leq h \leq M$,

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = m_h + \boldsymbol{\delta}^T \mathbf{z}_{t+h}, \quad V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = \begin{cases} \sigma^2_{t+1}, & h = 1 \\ v_h + \sigma^2_{t+h}, & h > 1 \end{cases}$$

Model	m_h	v_h
N-M	$\mu_t \cdot s_{t+h-M}$	$\sum_{i=0}^{h-2} \alpha_1^2 \cdot \sigma^2_{t+h-i} \cdot s^2_{t+h-M}/s^2_{t+h-M-i}$
A-M	$(\mu_t + h\beta_t) \cdot s_{t+h-M}$	$\sum_{i=0}^{h-2} [\alpha_1 + \alpha_2(i+1)]^2 \cdot \sigma^2_{t+h-i} \cdot s^2_{t+h-M}/s^2_{t+h-M-i}$
DA-M	$(\mu_t + \beta_t \sum_{i=1}^{h} \phi^i) \cdot s_{t+h-M}$	$\sum_{i=0}^{h-2} [\alpha_1 + \alpha_2 \sum_{j=1}^{i+1} \phi^j]^2 \cdot \sigma^2_{t+h-i} \cdot s^2_{t+h-M}/s^2_{t+h-M-i}$

The mapping $u(\boldsymbol{\beta}_t)$ could be any linear function of $\boldsymbol{\beta}_t$, and we consider only the case where $u(\boldsymbol{\beta}_t) = \mathbf{x}^T\boldsymbol{\beta}_t$. Similar to class 1, class 3 also consists of six models for N-N, A-N, DA-N, N-A, A-A, and DA-A respectively. The difference is that now the error depends on $\boldsymbol{\beta}_t$.

Derived in Appendix C, the conditional mean and variance of Y_{t+h} given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ for $h \geq 1$ are

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mathbf{x}^T\mathbf{G}^{h-1}\boldsymbol{\beta}_t + \boldsymbol{\delta}^T\mathbf{z}_{t+h}, \tag{4.37a}$$

$$V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = (1 + \sigma^2_{t+h})\mathbf{x}^T\mathbf{V}_{h-1}\mathbf{x} + \sigma^2_{t+h}(\mathbf{x}^T\mathbf{G}^{h-1}\boldsymbol{\beta}_t)^2, \tag{4.37b}$$

where

$$\mathbf{V}_h = \mathbf{G}\mathbf{V}_{h-1}\mathbf{G}^T + \sigma^2_{t+h}\boldsymbol{\alpha}\mathbf{x}^T\mathbf{V}_{h-1}\mathbf{x}\boldsymbol{\alpha}^T + \sigma^2_{t+h}(\mathbf{x}^T\mathbf{G}^{h-1}\boldsymbol{\beta}_t)^2\boldsymbol{\alpha}\boldsymbol{\alpha}^T, \tag{4.38a}$$

$$\mathbf{V}_0 = O, \tag{4.38b}$$

and O is a square matrix of zeros. The conditional means for models in class 3 are the same as those for the corresponding models in class 1. However, the condition variances for class 3 depend on $\boldsymbol{\beta}_t$ while the conditional variances for class 1 do not. This implies that corresponding models from class 1 and class 3 produce same point forecasts while different prediction intervals.

The simplest case in class 3 is the model for N-N. For this model,

$$\mathbf{x} = 1, \quad \boldsymbol{\beta}_t = \mu_t, \quad \mathbf{G} = 1, \quad \boldsymbol{\alpha} = \alpha_1. \tag{4.39}$$

With an additional assumption that ϵ_t has a constant variance, namely $E[\epsilon_t^2] = \sigma^2$, we have

$$\begin{aligned} \mathbf{V}_h &= (1 + \alpha_1^2\sigma_{t+h}^2)\mathbf{V}_{h-1} + \alpha_1^2\sigma_{t+h}^2\mu_t^2 \\ &= \alpha_1^2\sigma^2 \sum_{i=0}^{h-1}(1 + \alpha_1^2\sigma^2)^i\mu_t^2 = (1 + \alpha_1^2\sigma^2)^h\mu_t^2 - \mu_t^2. \end{aligned} \tag{4.40}$$

Therefore,

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mu_t + \boldsymbol{\delta}^T\mathbf{z}_{t+h}, \tag{4.41a}$$

$$V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = (1 + \sigma^2)(1 + \alpha_1^2\sigma^2)^{h-1}\mu_t^2 - \mu_t^2. \tag{4.41b}$$

While the model for N-N in class 1 has

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mu_t + \boldsymbol{\delta}^T\mathbf{z}_{t+h}, \tag{4.42a}$$

$$V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = (1 + (h-1)\alpha_1^2)\sigma^2. \tag{4.42b}$$

That is, the model for N-N in class 3 has the same condition mean as but different conditional variance than the model for N-N in class 1.

• Class 4 – Nonlinear SSOE model with multiplicative error

Replacing ϵ_t in model (4.34) by $u(\boldsymbol{\beta}_{t-1}^*, s_{t-M})\epsilon_t$ yields a model of the form

$$Y_t = \mathbf{x}^{*T}\boldsymbol{\beta}_{t-1}^* \cdot s_{t-M} + \boldsymbol{\delta}^T\mathbf{z}_t + u(\boldsymbol{\beta}_{t-1}^*, s_{t-M})\epsilon_t, \tag{4.43a}$$

$$\boldsymbol{\beta}_t^* = \mathbf{G}^*\boldsymbol{\beta}_{t-1}^* + \boldsymbol{\alpha}^* \cdot u(\boldsymbol{\beta}_{t-1}^*, s_{t-M})\epsilon_t / s_{t-M}, \tag{4.43b}$$

$$s_t = s_{t-M} + \alpha_3 \cdot u(\boldsymbol{\beta}_{t-1}^*, s_{t-M})\epsilon_t / (\mathbf{x}^{*T}\boldsymbol{\beta}_{t-1}^*). \tag{4.43c}$$

where $u(\boldsymbol{\beta}_{t-1}^*, s_{t-M})$ could be any mapping to $\Re$. We consider only the case where

$$u(\boldsymbol{\beta}_{t-1}^*, s_{t-M}) = \mathbf{x}^{*T}\boldsymbol{\beta}_{t-1}^* \cdot s_{t-M}. \tag{4.44}$$

This leads to a simpler form

$$Y_t = \mathbf{x}^{*T}\boldsymbol{\beta}^*_{t-1} \cdot s_{t-M} + \boldsymbol{\delta}^T\mathbf{z}_t + \mathbf{x}^{*T}\boldsymbol{\beta}^*_{t-1} \cdot s_{t-M} \cdot \epsilon_t, \tag{4.45a}$$

$$\boldsymbol{\beta}^*_t = \mathbf{G}^*\boldsymbol{\beta}^*_{t-1} + \boldsymbol{\alpha}^* \cdot \mathbf{x}^{*T}\boldsymbol{\beta}^*_{t-1} \cdot \epsilon_t, \tag{4.45b}$$

$$s_t = s_{t-M} + \alpha_3 \cdot s_{t-M} \cdot \epsilon_t. \tag{4.45c}$$

Class 4 contains three models that underpin ESCov with the intercept estimated by N-M, A-M, and DA-M respectively.

Derived in Appendix D, the condition mean and variance of Y_{t+h} given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ for $1 \le h \le M$ are

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mathbf{x}^{*T}(\mathbf{G}^*)^{h-1}\boldsymbol{\beta}^*_t \cdot s_{t+h-M} + \boldsymbol{\delta}^T\mathbf{z}_{t+h}, \tag{4.46a}$$

$$V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = s^2_{t+h-M}\left[(1+\sigma^2_{t+h})\mathbf{x}^{*T}\mathbf{V}^*_{h-1}\mathbf{x}^* + \sigma^2_{t+h}(\mathbf{x}^{*T}(\mathbf{G}^*)^{h-1}\boldsymbol{\beta}^*_t)^2\right], \tag{4.46b}$$

where

$$\mathbf{V}^*_h = \mathbf{G}^*\mathbf{V}^*_{h-1}\mathbf{G}^{*T} + \sigma^2_{t+h}\boldsymbol{\alpha}^*\mathbf{x}^{*T}\mathbf{V}^*_{h-1}\mathbf{x}^*\boldsymbol{\alpha}^{*T} + \sigma^2_{t+h}(\mathbf{x}^{*T}(\mathbf{G}^*)^{h-1}\boldsymbol{\beta}^*_t)^2\boldsymbol{\alpha}^*\boldsymbol{\alpha}^{*T},$$

$$\mathbf{V}^*_0 = O. \tag{4.47}$$

Comparing the results with those for class 2 shows that models for class 4 produce same point forecasts as but different prediction intervals from the corresponding models in class 2.

The simplest case in class 4 is the model for N-M. For this model,

$$\mathbf{x}^* = 1, \quad \boldsymbol{\beta}^*_t = \mu_t, \quad \mathbf{G}^* = 1, \quad \boldsymbol{\alpha}^* = \alpha_1, \tag{4.48}$$

which are the same as $\mathbf{x}$, $\boldsymbol{\beta}_t$, $\mathbf{G}$, and $\boldsymbol{\alpha}$ in the model underlying N-N in class 3. With the assumption that ϵ_t has a constant variance, namely $E[\epsilon^2_t] = \sigma^2$, we have, for $1 \le h \le M$,

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mu_t \cdot s_{t+h-M} + \boldsymbol{\delta}^T\mathbf{z}_{t+h}, \tag{4.49a}$$

$$V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = s^2_{t+h-M} \cdot [(1+\sigma^2)(1+\alpha^2_1\sigma^2)^{h-1}\mu^2_t - \mu^2_t]. \tag{4.49b}$$

While the model for N-M in class 2 has

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mu_t \cdot s_{t+h-M} + \boldsymbol{\delta}^T \mathbf{z}_{t+h}, \tag{4.50a}$$

$$V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) = s^2_{t+h-M} \cdot \alpha_1^2 \cdot \sigma^2 \cdot \sum_{i=0}^{h-2} \frac{1}{s^2_{t+h-M-i}}. \tag{4.50b}$$

That is, the two models for N-M have same conditional means but different conditional variances.

4.5 Related Statistical Model

Comparing SSOE state space models underlying ESCov in Table 4.3 with those underlying ES methods (see Table 2.4 in Chapter 2) reveals that the former is different from the latter in that an extra term, $\boldsymbol{\delta}^T \mathbf{z}_t$, is introduced into the observation equation. In Chapter 2, we showed that SSOE models underlying additive ES methods can be reduced to an ARIMA form. In a similar fashion, it can be showed that SSOE models underlying additive ESCov's also have a reduced form, and moreover the reduced form is a transfer function model, an extension of the ARIMA model (Box, Jenkins and Reinse 1994).

For example, the SSOE model underlying SES has the transition equation

$$\mu_t = \mu_{t-1} + \alpha_1 \epsilon_t, \tag{4.51}$$

which can be rewritten as

$$(1 - B)\mu_t = \alpha_1 \epsilon_t. \tag{4.52}$$

Left multiplying the observation equation by $1 - B$ and substituting equation (4.52) into the result gives

$$(1 - B)Y_t = \boldsymbol{\delta}^T (1 - B)\mathbf{z}_t + \epsilon_t + (\alpha_1 - 1)\epsilon_{t-1}. \tag{4.53}$$

This is a transfer function model with input $\mathbf{z}_t$ and output Y_t. Table 4.8 lists the reduced form transfer function models of SSOE state space models that underpin additive ESCov's.

Table 4.8: Reduced Form Transfer Function Models of SSOE State Space Underlying Additive ESCov, (**N** - None, **A** - Additive, **DA** - Damped Additive

	Seasonality	
Trend	**N**	**A**
N	$(1-B)Y_t = \boldsymbol{\delta}^T(1-B)\mathbf{z}_t + (1+\theta_1 B)\epsilon_t$ $\theta_1 = \alpha_1 - 1$	$(1-B^M)Y_t = \boldsymbol{\delta}^T(1-B^M)\mathbf{z}_t + (1+\sum_{i=1}^{M}\theta_i B^i)\epsilon_t$ $\theta_i = \alpha_1, \quad i = 1, \cdots, M-1$ $\theta_M = \alpha_1 + \alpha_3 - 1$
A	$(1-B)^2 Y_t$ $= \boldsymbol{\delta}^T(1-B)^2\mathbf{z}_t + (1+\theta_1 B + \theta_2 B^2)\epsilon_t$ $\theta_1 = \alpha_1 + \alpha_2 - 2$ $\theta_2 = 1 - \alpha_1$	$(1-B)(1-B^M)Y_t$ $= \boldsymbol{\delta}^T(1-B)(1-B^M)\mathbf{z}_t + (1+\sum_{i=1}^{M+1}\theta_i B^i)\epsilon_t$ $\theta_1 = \alpha_1 + \alpha_2 - 1$ $\theta_i = \alpha_2, \quad i = 2, \cdots, M-1$ $\theta_M = \alpha_2 + \alpha_3 - 1$ $\theta_{M+1} = 1 - \alpha_1 - \alpha_3$
DA	$(1-\phi B)(1-B)Y_t$ $= \boldsymbol{\delta}^T(1-\phi B)(1-B)\mathbf{z}_t + (1+\theta_1 B + \theta_2 B^2)\epsilon_t$ $\theta_1 = \alpha_1 + \phi\alpha_2 - \phi - 1$ $\theta_2 = \phi(1-\alpha_1)$	$(1-\phi B)(1-B^M)Y_t$ $= \boldsymbol{\delta}^T(1-\phi B)(1-B^M)\mathbf{z}_t + (1+\sum_{i=1}^{M+1}\theta_i B^i)\epsilon_t$ $\theta_1 = \alpha_1 + \phi\alpha_2 - \phi$ $\theta_i = (1-\phi)\alpha_1 + \phi\alpha_2, \quad i = 2, \cdots, M-1$ $\theta_M = (1-\phi)\alpha_1 + \phi\alpha_2 + \alpha_3 - 1$ $\theta_{M+1} = \phi(1-\alpha_1) - \phi\alpha_3$

4.6 Appendix

A. Derivation of $E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}]$ and $V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h})$ for Models in Class 1

From equation (4.32b), we have

$$\boldsymbol{\beta}_{t+h} = \mathbf{G}\boldsymbol{\beta}_{t+h-1} + \boldsymbol{\alpha}\epsilon_{t+h} = \mathbf{G}^h\boldsymbol{\beta}_t + \sum_{i=0}^{h-1} \mathbf{G}^i\boldsymbol{\alpha}\epsilon_{t+h-i}, \tag{4.54}$$

which leads to the conditional mean and variance of $\boldsymbol{\beta}_{t+h}$ given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$

$$E[\boldsymbol{\beta}_{t+h}|\boldsymbol{\beta}_t] = \mathbf{G}^h\boldsymbol{\beta}_t, \tag{4.55a}$$

$$V(\boldsymbol{\beta}_{t+h}|\boldsymbol{\beta}_t) = \sum_{i=0}^{h-1} \mathbf{G}^i\boldsymbol{\alpha}\boldsymbol{\alpha}^T(\mathbf{G}^i)^T\sigma^2_{t+h-i}. \tag{4.55b}$$

Since, according to equation (4.32a),

$$Y_{t+h} = \mathbf{x}^T\boldsymbol{\beta}_{t+h-1} + \boldsymbol{\delta}^T\mathbf{z}_{t+h} + \epsilon_{t+h}, \tag{4.56}$$

we have the conditional mean and variance of Y_{t+h} given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ as

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mathbf{x}^T E[\boldsymbol{\beta}_{t+h-1}|\boldsymbol{\beta}_t] + \boldsymbol{\delta}^T\mathbf{z}_{t+h} = \mathbf{x}^T\mathbf{G}^{h-1}\boldsymbol{\beta}_t + \boldsymbol{\delta}^T\mathbf{z}_{t+h} \tag{4.57}$$

and

$$\begin{aligned} V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) &= \mathbf{x}^T V(\boldsymbol{\beta}_{t+h-1}|\boldsymbol{\beta}_t)\mathbf{x} + \sigma^2_{t+h} \\ &= \begin{cases} \sigma^2_{t+1} & h = 1 \\ \sum_{i=0}^{h-2}(\mathbf{x}^T\mathbf{G}^i\boldsymbol{\alpha})^2 \cdot \sigma^2_{t+h-i} + \sigma^2_{t+h} & h > 1 \end{cases} \end{aligned} \tag{4.58}$$

To get the specific results for each of the six models in class 1, we start with the model for DA-A as the other five can be considered as its special cases. The model for DA-A has, as given in Table 4.5,

$$\mathbf{x} = (1,\ \phi,\ \mathbf{0}'_{M-1}, 1)^T, \quad \boldsymbol{\beta}_t = (\mu_t,\ \beta_t,\ s_t,\ \cdots,\ s_{t-M+1})^T,$$

$$\mathbf{G} = \begin{bmatrix} 1 & \phi & \mathbf{0}'_{M-1} & 0 \\ 0 & \phi & \mathbf{0}'_{M-1} & 0 \\ 0 & 0 & \mathbf{0}'_{M-1} & 1 \\ \mathbf{0}_{M-1} & \mathbf{0}_{M-1} & \mathbf{I}_{M-1} & \mathbf{0}_{M-1} \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \mathbf{0}'_{M-1} \end{bmatrix}. \tag{4.59}$$

As

$$\mathbf{G}^k = \begin{bmatrix} 1 & \sum_{i=1}^{k} \phi^i & \mathbf{0}'_{M-k^*} & \mathbf{0}_{k^*} \\ 0 & \phi^k & \mathbf{0}'_{M-k^*} & \mathbf{0}_{k^*} \\ \mathbf{0}_{k^*} & \mathbf{0}_{k^*} & \mathbf{0}_{k^*} \times \mathbf{0}'_{M-k^*} & \mathbf{I}_{k^*} \\ \mathbf{0}_{M-k^*} & \mathbf{0}_{M-k^*} & \mathbf{I}_{M-k^*} & \mathbf{0}_{M-k^*} \times \mathbf{0}'_{k^*} \end{bmatrix}, \tag{4.60}$$

where $k^* = k(\text{mod } M)$, we have

$$\mathbf{x}^T\mathbf{G}^k = (1,\ \sum_{i=1}^{k+1} \phi^i,\ \mathbf{0}'_{M-k^*-1},\ 1,\ \mathbf{0}'_{k^*}). \tag{4.61}$$

Therefore,

$$m_h = \mathbf{x}^T\mathbf{G}^{h-1}\boldsymbol{\beta}_t = \mu_t + \beta_t \sum_{i=1}^{h} \phi^i + s_{t-M+h^*} \tag{4.62a}$$

$$\begin{aligned} v_h &= \sum_{i=0}^{h-2} (\mathbf{x}^T\mathbf{G}^i\boldsymbol{\alpha})^2 \cdot \sigma^2_{t+h-i} \\ &= \sum_{i=0}^{h-2} \left[\alpha_1 + \alpha_2 \cdot \sum_{j=1}^{i+1} \phi^j + \alpha_3 \cdot I_{i,M-1}\right]^2 \sigma^2_{t+h-i}, \quad h > 1 \end{aligned} \tag{4.62b}$$

where $I_{i,M-1} = 1$, if $i^* = M-1$; $= 0$, otherwise.

m_h and v_h for the other five models in class 1 can be derived from equations (4.62a) and (4.62b) by letting

(1). $\phi = 0$; $s_t = 0, \forall t$; and $\alpha_3 = 0$ for N-N

(2). $\phi = 1$; $s_t = 0, \forall t$; and $\alpha_3 = 0$ for A-N

(3). $s_t = 0, \forall t$; and $\alpha_3 = 0$ for DA-N

(4). $\phi = 0$ for N-A

(5). $\phi = 1$ for A-A

B. Derivation of $E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}]$ and $V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h})$ for Models in Class 2

According to equation (4.34b),

$$\boldsymbol{\beta}^*_{t+h} = \mathbf{G}^*\boldsymbol{\beta}^*_{t+h-1} + \boldsymbol{\alpha}^*\epsilon_{t+h}/s_{t+h-M} = (\mathbf{G}^*)^h\boldsymbol{\beta}^*_t + \sum_{i=0}^{h-1} (\mathbf{G}^*)^i\boldsymbol{\alpha}^* \cdot \epsilon_{t+h-i}/s_{t+h-M-i}. \tag{4.63}$$

Then, for $1 \leq h \leq M$, the conditional mean and variance of $\boldsymbol{\beta}^*_{t+h}$ given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ are

$$E[\boldsymbol{\beta}^*_{t+h}|\boldsymbol{\beta}_t] = (\mathbf{G}^*)^h \boldsymbol{\beta}^*_t, \tag{4.64}$$

and

$$V(\boldsymbol{\beta}^*_{t+h}|\boldsymbol{\beta}_t) = \sum_{i=0}^{h-1} (\mathbf{G}^*)^i \boldsymbol{\alpha}^* \boldsymbol{\alpha}^{*T} ((\mathbf{G}^*)^i)^T \cdot \sigma^2_{t+h-i}/s^2_{t+h-M-i}. \tag{4.65}$$

According to equation (4.34a),

$$Y_{t+h} = \mathbf{x}^{*T} \boldsymbol{\beta}^*_{t+h-1} \cdot s_{t+h-M} + \boldsymbol{\delta}^T \mathbf{z}_{t+h} + \epsilon_{t+h}. \tag{4.66}$$

Hence, for $1 \leq h \leq M$, the conditional mean and variance of Y_{t+h} given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ are

$$\begin{aligned} E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] &= \mathbf{x}^{*T} E[\boldsymbol{\beta}^*_{t+h-1}|\boldsymbol{\beta}_t] \cdot s_{t+h-M} + \boldsymbol{\delta}^T \mathbf{z}_{t+h} \\ &= \mathbf{x}^{*T} (\mathbf{G}^*)^{h-1} \boldsymbol{\beta}^*_t \cdot s_{t+h-M} + \boldsymbol{\delta}^T \mathbf{z}_{t+h}, \end{aligned} \tag{4.67}$$

and

$$\begin{aligned} & V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) \\ &= \mathbf{x}^{*T} V(\boldsymbol{\beta}^*_{t+h-1}|\boldsymbol{\beta}_t) \mathbf{x}^* \cdot s^2_{t+h-M} + \sigma^2_{t+h} \\ &= \begin{cases} \sigma^2_{t+1}, & h = 1 \\ \sum_{i=0}^{h-2} (\mathbf{x}^{*T} (\mathbf{G}^*)^i \boldsymbol{\alpha}^*)^2 \cdot \sigma^2_{t+h-i} \cdot s^2_{t+h-M}/s^2_{t+h-M-i} + \sigma^2_{t+h}, & h > 1 \end{cases} \end{aligned} \tag{4.68}$$

Similarly, to get the results for a particular model in class 2, we start with the general one, the model for DA-M. For this model,

$$\mathbf{x}^* = \begin{bmatrix} 1 \\ \phi \end{bmatrix}, \quad \boldsymbol{\beta}^*_t = \begin{bmatrix} \mu_t \\ \beta_t \end{bmatrix}, \quad \mathbf{G}^* = \begin{bmatrix} 1 & \phi \\ 0 & \phi \end{bmatrix}, \quad \boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}. \tag{4.69}$$

Since

$$(\mathbf{G}^*)^k = \begin{bmatrix} 1 & \sum_{i=1}^k \phi^i \\ 0 & \phi^k \end{bmatrix}, \tag{4.70}$$

we have

$$\mathbf{x}^{*T}(\mathbf{G}^*)^k = (1,\ \sum_{i=1}^{k+1}\phi^i). \tag{4.71}$$

Therefore,

$$\begin{aligned} m_h &= \mathbf{x}^{*T}(\mathbf{G}^*)^{h-1}\boldsymbol{\beta}_t^* \cdot s_{t+h-M} = (\mu_t + \beta_t \sum_{i=1}^{h}\phi^i)\cdot s_{t+h-M}, && \text{(4.72a)} \\ v_h &= \sum_{i=0}^{h-2}(\mathbf{x}^{*T}\mathbf{G}^{*i}\boldsymbol{\alpha}^*)^2 \cdot \sigma^2_{t+h-i}\cdot s^2_{t+h-M}/s^2_{t+h-M-i} \\ &= \sum_{i=0}^{h-2}\left[\alpha_1 + \alpha_2 \cdot \sum_{j=1}^{i+1}\phi^j\right]^2 \cdot \sigma^2_{t+h-i}\cdot s^2_{t+h-M}/s^2_{t+h-M-i}, \quad h > 1 && \text{(4.72b)} \end{aligned}$$

Models for N-M and DA-M are the special cases of the model for DA-M, and the m_h and v_h for them can be derived from equations (4.72a) and (4.72b) by letting $\phi = 0$ for N-M and $\phi = 1$ for A-M.

C. Derivation of $E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}]$ and $V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h})$ for Models in Class 3

Let $\mathbf{M}_h = E[\boldsymbol{\beta}_{t+h}|\boldsymbol{\beta}_t]$ and $\mathbf{V}_h = V(\boldsymbol{\beta}_{t+h}|\boldsymbol{\beta}_t)$. Then, $\mathbf{M}_0 = \boldsymbol{\beta}_t$ and $\mathbf{V}_0 = O$, where O is a square matrix of zeros. According to equation (4.36b),

$$\boldsymbol{\beta}_{t+h} = \mathbf{G}\boldsymbol{\beta}_{t+h-1} + \boldsymbol{\alpha}\cdot\mathbf{x}^T\boldsymbol{\beta}_{t+h-1}\cdot\epsilon_{t+h}, \tag{4.73}$$

Therefore, the conditional mean and variance of $\boldsymbol{\beta}_{t+h}$ given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ are

$$\begin{aligned} \mathbf{M}_h &= \mathbf{G}\mathbf{M}_{h-1} = \mathbf{G}^h\boldsymbol{\beta}_t, && \text{(4.74a)} \\ \mathbf{V}_h &= \mathbf{G}\mathbf{V}_{h-1}\mathbf{G}^T + \boldsymbol{\alpha}\mathbf{x}^T V(\boldsymbol{\beta}_{t+h-1}\cdot\epsilon_{t+h}|\boldsymbol{\beta}_t)\mathbf{x}\boldsymbol{\alpha}^T \\ &= \mathbf{G}\mathbf{V}_{h-1}\mathbf{G}^T + \boldsymbol{\alpha}\mathbf{x}^T E[\boldsymbol{\beta}_{t+h-1}\boldsymbol{\beta}_{t+h-1}^T|\boldsymbol{\beta}_t]\cdot E[\epsilon^2_{t+h}]\mathbf{x}\boldsymbol{\alpha}^T \\ &= \mathbf{G}\mathbf{V}_{h-1}\mathbf{G}^T + \sigma^2_{t+h}\boldsymbol{\alpha}\mathbf{x}^T(\mathbf{V}_{h-1} + \mathbf{M}_{h-1}\mathbf{M}_{h-1}^T)\mathbf{x}\boldsymbol{\alpha}^T \\ &= \mathbf{G}\mathbf{V}_{h-1}\mathbf{G}^T + \sigma^2_{t+h}\boldsymbol{\alpha}\mathbf{x}^T\mathbf{V}_{h-1}\mathbf{x}\boldsymbol{\alpha}^T + \sigma^2_{t+h}(\mathbf{x}^T\mathbf{G}^{h-1}\boldsymbol{\beta}_t)^2\boldsymbol{\alpha}\boldsymbol{\alpha}^T && \text{(4.74b)} \end{aligned}$$

Since, according to equation (4.36a),

$$Y_{t+h} = \mathbf{x}^T\boldsymbol{\beta}_{t+h-1} + \boldsymbol{\delta}^T\mathbf{z}_{t+h} + \mathbf{x}^T\boldsymbol{\beta}_{t+h-1}\cdot\epsilon_{t+h}, \tag{4.75}$$

we have the conditional mean and variance of Y_{t+h} given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ as

$$E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] = \mathbf{x}^T\mathbf{M}_{h-1} + \boldsymbol{\delta}^T\mathbf{z}_{t+h} = \mathbf{x}^T\mathbf{G}^{h-1}\boldsymbol{\beta}_t + \boldsymbol{\delta}^T\mathbf{z}_{t+h} \tag{4.76}$$

and

$$\begin{aligned} V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) &= \mathbf{x}^T\mathbf{V}_{h-1}\mathbf{x} + \mathbf{x}^T V(\boldsymbol{\beta}_{t+h-1}\cdot\epsilon_{t+h}|\boldsymbol{\beta}_t)\mathbf{x} \\ &= (1+\sigma^2_{t+h})\mathbf{x}^T\mathbf{V}_{h-1}\mathbf{x} + \sigma^2_{t+h}(\mathbf{x}^T\mathbf{G}^{h-1}\boldsymbol{\beta}_t)^2. \end{aligned} \tag{4.77}$$

D. Derivation of $E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}]$ and $V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h})$ for Models in Class 4

Let $\mathbf{M}^*_h = E[\boldsymbol{\beta}^*_{t+h}|\boldsymbol{\beta}_t]$ and $\mathbf{V}^*_h = V(\boldsymbol{\beta}^*_{t+h}|\boldsymbol{\beta}_t)$. Then, $\mathbf{M}^*_0 = \boldsymbol{\beta}^*_t$ and $\mathbf{V}^*_0 = O$. According to equation (4.45b),

$$\boldsymbol{\beta}^*_{t+h} = \mathbf{G}^*\boldsymbol{\beta}^*_{t+h-1} + \boldsymbol{\alpha}^*\cdot\mathbf{x}^{*T}\boldsymbol{\beta}^*_{t+h-1}\cdot\epsilon_{t+h}, \tag{4.78}$$

which looks the same as equation (4.73). Following the same line of reasoning, we obtain the conditional mean and variance of $\boldsymbol{\beta}_{t+h}$ given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$

$$\mathbf{M}^*_h = (\mathbf{G}^*)^h\boldsymbol{\beta}^*_t, \tag{4.79a}$$

$$\mathbf{V}^*_h = \mathbf{G}^*\mathbf{V}^*_{h-1}\mathbf{G}^{*T} + \sigma^2_{t+h}\boldsymbol{\alpha}^*\mathbf{x}^{*T}\mathbf{V}^*_{h-1}\mathbf{x}^*\boldsymbol{\alpha}^{*T} + \sigma^2_{t+h}(\mathbf{x}^{*T}(\mathbf{G}^*)^{h-1}\boldsymbol{\beta}^*_t)^2\boldsymbol{\alpha}^*\boldsymbol{\alpha}^{*T}. \tag{4.79b}$$

Since, according to equation (4.45a),

$$Y_{t+h} = \mathbf{x}^{*T}\boldsymbol{\beta}^*_{t+h-1}\cdot s_{t+h-M} + \boldsymbol{\delta}^T\mathbf{z}_{t+h} + \mathbf{x}^{*T}\boldsymbol{\beta}^*_{t+h-1}\cdot s_{t+h-M}\cdot\epsilon_{t+h}, \tag{4.80}$$

we have the conditional mean and variance of Y_{t+h} given $\boldsymbol{\beta}_t$ and $\mathbf{z}_{t+h}$ for $1 \le h \le M$ as

$$\begin{aligned} E[Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}] &= \mathbf{x}^{*T}\mathbf{M}^*_{h-1}\cdot s_{t+h-M} + \boldsymbol{\delta}^T\mathbf{z}_{t+h} \\ &= \mathbf{x}^{*T}(\mathbf{G}^*)^{h-1}\boldsymbol{\beta}^*_t\cdot s_{t+h-M} + \boldsymbol{\delta}^T\mathbf{z}_{t+h} \end{aligned} \tag{4.81}$$

and

$$
\begin{aligned}
& V(Y_{t+h}|\boldsymbol{\beta}_t, \mathbf{z}_{t+h}) && (4.82)\\
&= \mathbf{x}^{*T}\mathbf{V}^*_{h-1}\mathbf{x}^* \cdot s^2_{t+h-M} + \mathbf{x}^{*T} V(\boldsymbol{\beta}^*_{t+h-1} \cdot \epsilon_{t+h}|\boldsymbol{\beta}_t)\mathbf{x}^* \cdot s^2_{t+h-M} \\
&= s^2_{t+h-M}(1+\sigma^2_{t+h})\mathbf{x}^{*T}\mathbf{V}^*_{h-1}\mathbf{x}^* + s^2_{t+h-M}\sigma^2_{t+h}(\mathbf{x}^{*T}(\mathbf{G}^*)^{h-1}\boldsymbol{\beta}^*_t)^2. && (4.83)
\end{aligned}
$$

CHAPTER V

BAYESIAN VALIDATION OF COMPUTER MODELS

5.1 Introduction

Computer models are mathematic representations of real systems, for example a group of partial differential equations with initial and boundary conditions for many engineering problems. They have been commonly used to investigate complex systems for which physical experiments are either highly expensive or too time-consuming (Sacks et al 1989, Welch et al 1992, Santner et al 2003). However, before using a computer model to investigate a real system, we need to address an important question "How well does the computer model represent the real system?" Without a meaningful answer to this question, any conclusions based on the analysis of outputs from a computer model are about this computer model and can not be simply applied to the real system of interest. The process of determining to what degree a computer model accurately represents the real system is referred to as model validation (AIAA G-077-1998) that generally involves the comparison of outputs computed from a computer model to observations collected from physical experiments.

There are many ways to compare computer outputs and physical observations for validating a computer model. For example, we can graphically display both computer outputs and physical observations in one plot and see if computer outputs agree with physical observations. The graphical comparison is easy and simple and probably is the first thing we should do before attempting any sophisticated methods. However,

this approach is obviously too subjective and lacks a quantitative indication of agreement between computer outputs and physical observation. A more statistically sound way for the comparison of computer outputs and physical observations is to conduct a hypothesis testing. Hypothesis testing is not a new technique, but not until recently did people start to use it for computer model validation (Hills and Trucano 1999, Hills and Trucano 2002, Hills 2006). For example, Hills and Trucano (2002) used a χ^2 test for computer model validation. They assumed that the vectors of computer outputs and physical observations follow independent multivariate normal distributions and computed a χ^2 statistic to test a null hypothesis that the model bias (i.e., the difference of the two vectors) has a zero mean. Using hypothesis testing, we can control the type-I error (i.e., the probability of rejecting a good computer model) but not type-II error (i.e., the probability of failing to reject a bad computer model). In other words, rejecting the null hypothesis means that we have strong evidence that the computer model is not an accurate representation of the real system. However, failing to reject a null hypothesis could be due to two reasons, the computer model does not accurately represent the real system or we do not have enough data to reject a bad computer model. Furthermore, hypothesis testing, after rejecting a computer model, gives no suggestions on how the model bias behaves and how the prediction by the computer model can be improved with available physical observations.

Recently, Oberkampf and Barone (2004) gave a comprehensive review on computer model validation. They argued that computer model validation should be done quantitatively through the use of computable measures that enable the quantitative comparison of computer outputs and physical observations over a range of input variables. Those measures have been referred to as validation metrics. Oberkampf and Barone (2004) discussed a variety of conceptual properties that a validation metric should possess and emphasized that a validation metric should quantify uncertainties in the comparison of computer outputs and physical observations. Uncertainties could

be due to random measurement errors in the physical observations and/or errors resulting from post-processing computer outputs and/or physical observations, such as errors resulting from fitting a model to computer outputs or physical observations. In the same paper, they proposed a validation metric that uses the statistical concept of confidence intervals for the quantification of uncertainties. With computer outputs and physical observations as the input data, they first fitted a nonlinear regression model to physical observations and constructed a confident band for the fitted regression curve. This confidence band is composed of individual confidence intervals for the fitted curve at different values of input variables. They then compared the fitted regression curve along with its confidence band to computer outputs. This comparison should be able to reveal how well computer outputs agree with physical observations and in which regions of the input space the two agree with each other well and in which regions they do not. Oberkampf and Barone (2004) for the first time provide an approach that validates a computer model by quantitatively comparing computer outputs and physical observations over a range of input variables.

However, there are some concerns with Oberkampf and Barone's approach. First, the choice of the nonlinear regression model has a great influence in validation results. A large disagreement between the fitted regression curve and computer outputs might be due to an inappropriate choice of the regression model rather than a poor computer model. An appropriate model choice requires good scientific knowledge of the real system being studied. Second, the fact that often only few physical observations are available implies likely difficulties in the fitting of the nonlinear regression model as nonlinear models usually require a large number of observations to have a good model estimation. Furthermore, the model fitting uses only physical observations and does not consider computer outputs. We will show later that integrating together physical observations and computer outputs produces a better prediction of the real system. Third, Oberkampf and Barone (2004) quantified uncertainties due to measurement

errors in physical observations and estimation errors in model fitting by constructing confidence intervals for the fitted regression curve. However, with nonlinear models, the computation of confidence intervals is rather complicated and often requires certain approximations. Fourth, with Oberkampf and Barone's approach, it is not clear how to improve the prediction of the real system when the comparison suggests a large disagreement between computation and experimentation.

The validation method proposed by Oberkampf and Barone (2004) is a frequentist approach. In this chapter, we proposed a Bayesian approach to computer model validation. The Bayesian approach has the ability to take into consideration prior knowledge on the real system in the form of prior distributions for certain parameters. The outputs of our Bayesian approach include the posterior distributions of both the model bias and the real system output. The posterior distribution of the model bias serves as a validation metric for the quantitative comparison of computer outputs and physical observations. Both the mean and variance of the model bias are functions of input variables, providing an evaluation of the representativeness of the computer model over a range of input variables.

The proposed Bayesian approach overcomes certain problems Oberkampf and Barone's approach has. First, with the validation metric in the form of a probability distribution, the construction of confidence intervals is straightforward and requires no approximations. Second, the proposed Bayesian approach models the real system output as the sum of two Gaussian processes. Gaussian processes are essentially nonparametric and can adapt to any shape suggested by the data with only a simple assumption for the mean function, which frees us from choosing a complicated nonlinear model. Third, as most parameters are integrated out during the derivation of the posterior distributions, the estimation becomes much easier compared to fitting a nonlinear regression model. Fourth, as mentioned above, the proposed Bayesian approach also produces the posterior distribution of the real system output.

This posterior is based on both physical observations and computer outputs, giving a better prediction of the real system than using only either physical observations or computer outputs as demonstrated in section 5.6. Furthermore, the posterior of the real system output considers the uncertainties due to model fitting to computer outputs. This source of uncertainties were ignored by Oberkampf and Barone (2004).

In some aspects, our proposed Bayesian approach is similar to the Bayesian approach proposed by Kennedy and O'Hagan (2001) for the calibration of computer models using both computer outputs and physical observations. Their approach also uses Gaussian processes and assumes a similar relationship among physical observations, computer outputs, and the real system output, essentially equation (5.1) with the term $Y^m(\mathbf{x})$ replaced by $\rho Y^m(\mathbf{x}, \Theta)$, where ρ is an unknown constant, and Θ is the vector of calibration parameters. The presence of Θ is important, indicating that their method is aimed at finding the value of Θ that brings computer outputs as close as possible to physical observations rather than modelling the difference between them as we will do. In addition, Kennedy and O'Hagan (2001) assume improper priors for unknown parameters, which is equivalent to treating them as fixed unknowns. While, our approach adopts traditional priors–normal distributions for location parameters and inverse gamma distributions for variance parameters–and is a full Bayesian analysis as we integrate out both location and variance parameters.

The organization of this chapter is as follows. In section 5.2, we give the general statistical framework on which our Bayesian approach is based. In section 5.3, we derive the posterior distributions of the model bias, the computer output, and the real system output. In section 5.4, we discuss the estimation of certain parameters present in the derived posterior distributions. In section 5.5, based on the results presents in section 5.3 and 5.4, we describe a complete Bayesian procedure for computer model validation. In section 5.6, we demonstrate the proposed Bayesian approach using several numerical examples.

5.2 Statistical Framework

Let $Y^m(\mathbf{x})$ be the output of a computer model at $\mathbf{x}$, where $\mathbf{x} = (x_1, \cdots, x_p)^T$ is a point in a p-dimensional input space. Let $Y^e(\mathbf{x})$ and $Y^r(\mathbf{x})$ be the physical observation and the real system output at $\mathbf{x}$ respectively. The proposed Bayesian approach models the relationships among the computer model output $Y^m(\mathbf{x})$, the physical observation $Y^e(\mathbf{x})$, and the real system output $Y^r(\mathbf{x})$ via

$$Y^e(\mathbf{x}) = Y^r(\mathbf{x}) + \epsilon(\mathbf{x}) = Y^m(\mathbf{x}) + \delta(\mathbf{x}) + \epsilon(\mathbf{x}), \tag{5.1}$$

where $\delta(\mathbf{x})$ is the bias of the computer model, and $\epsilon(\mathbf{x})$ is the measurement error of the physical observation.

For the model in equation (5.1), we assume the following:

- The computer model output $Y^m(\mathbf{x})$ is a Gaussian process with mean $\mu_m(\mathbf{x}) = \mathbf{f}_m^T(\mathbf{x})\boldsymbol{\beta}_m$ and covariance function $\sigma_m^2 R_m$, where $\mathbf{f}_m(\mathbf{x}) = (f_{m,1}(\mathbf{x}), \cdots, f_{m,q_m}(\mathbf{x}))^T$ is a vector of q_m functions of $\mathbf{x}$, and R_m has an exponential form

$$R_m(\mathbf{x}_i, \mathbf{x}_j) = \prod_{k=1}^{p} \exp\left\{-\phi_{m,k}(x_{i,k} - x_{j,k})^{P_{m,k}}\right\}, \tag{5.2}$$

  with $\phi_{m,k} > 0$ and $0 < P_{m,k} \leq 2$ for $k = 1, \cdots, p$. We denote $\boldsymbol{\phi}_m = (\phi_{m,1}, \cdots, \phi_{m,p})^T$ and $\mathbf{P}_m = (P_{m,1}, \cdots, P_{m,p})^T$.

- The model bias $\delta(\mathbf{x})$ is a Gaussian process with mean $\mu_\delta(\mathbf{x}) = \mathbf{f}_\delta^T(\mathbf{x})\boldsymbol{\beta}_\delta$ and covariance function $\sigma_\delta^2 R_\delta$, where $\mathbf{f}_\delta(\mathbf{x}) = (f_{\delta,1}(\mathbf{x}), \cdots, f_{\delta,q_\delta}(\mathbf{x}))^T$ is a vector of q_δ functions of $\mathbf{x}$, and R_δ has an exponential form

$$R_\delta(\mathbf{x}_i, \mathbf{x}_j) = \prod_{k=1}^{p} \exp\left\{-\phi_{\delta,k}(x_{i,k} - x_{j,k})^{P_{\delta,k}}\right\}, \tag{5.3}$$

  with $\phi_{\delta,k} > 0$ and $0 < P_{\delta,k} \leq 2$ for $k = 1, \cdots, p$. We denote $\boldsymbol{\phi}_\delta = (\phi_{\delta,1}, \cdots, \phi_{\delta,p})^T$ and $\mathbf{P}_\delta = (P_{\delta,1}, \cdots, P_{\delta,p})^T$.

- The measurement error $\epsilon(\mathbf{x})$ has a normal distriution with mean zero and variance σ_ϵ^2, and $E[\epsilon(\mathbf{x}_i) \cdot \epsilon(\mathbf{x}_j)]$=0 for any $\mathbf{x}_i \neq \mathbf{x}_j$.

- $Y^m(\cdot)$, $\delta(\cdot)$, and $\epsilon(\cdot)$ are mutually independent.

Before continuing , we give a few notations. Let $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ and $D^* = \{\mathbf{x}_1^*, \cdots, \mathbf{x}_{n^*}^*\}$ be any two sets of points in the input space. Let $\mathbf{R}_m(D, D^*)$ be the matrix of correlations between the vectors $Y^m(D) = (Y^m(\mathbf{x}_1), \cdots, Y^m(\mathbf{x}_n))^T$ and $Y^m(D^*) = (Y^m(\mathbf{x}_1^*), \cdots, Y^m(\mathbf{x}_{n^*}^*))^T$ with the $(i,j)^{th}$ element $R_m(\mathbf{x}_i, \mathbf{x}_j^*)$, and use $\mathbf{R}_m(D)$ as a shorthand for $\mathbf{R}_m(D, D)$. Let $\mathbf{F}_m(D) = (\mathbf{f}_m(\mathbf{x}_1), \cdots, \mathbf{f}_m(\mathbf{x}_n))^T$. Similarly, we define $\mathbf{R}_\delta(D, D^*)$, $\mathbf{R}_\delta(D)$, and $\mathbf{F}_\delta(D)$. Let $D_e = \{\mathbf{x}_1^e, \cdots, \mathbf{x}_{n_e}^e\}$ be the set of n_e points in the input space where physical observations are available, and $D_m = \{\mathbf{x}_1^m, \cdots, \mathbf{x}_{n_m}^m\}$ the set of n_m points where computer outputs are available. D_e and D_m may or may not overlap. Let $\mathbf{y}^e = (y^e(\mathbf{x}_1^e), \cdots, y^e(\mathbf{x}_{n_e}^e))^T$ and $\mathbf{y}^m = (y^m(\mathbf{x}_1^m), \cdots, y^m(\mathbf{x}_{n_m}^m))^T$ be the vectors of physical observations at D_e and computer outputs at D_m respectively. Let $\mathbf{I}_k$ be an $k \times k$ identity matrix.

5.3 The Bayesian Approach

In this section, we derive the posterior distributions of the model bias $\delta(\cdot)$, the computer output $Y^m(\cdot)$, and the real system output $Y^r(\cdot)$ under two assumptions: (1) $D_e \subseteq D_m$, and (2) certain parameters such as $\boldsymbol{\phi}_\delta$ and $\mathbf{P}_\delta$ are known. In next section, we deal with situations where those two assumptions do not hold.

5.3.1 Prior Distributions for Unknown Parameters

One advantage of the proposed Bayesian approach is its ability to take into account a priori knowledge on the real system in the form of prior distributions for unknown parameters. Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \sigma_\epsilon^2\}$ be the collection of parameters in equation (5.1). We assume the following priors

$$\boldsymbol{\beta}_\delta | \sigma_\delta^2 \sim N(\mathbf{b}_\delta, \sigma_\delta^2 \mathbf{V}_\delta), \quad \boldsymbol{\beta}_m | \sigma_m^2 \sim N(\mathbf{b}_m, \sigma_m^2 \mathbf{V}_m), \tag{5.4a}$$

$$\sigma_\delta^2 \sim IG(\alpha_\delta, \gamma_\delta), \quad \sigma_m^2 \sim IG(\alpha_m, \gamma_m), \tag{5.4b}$$

where $N(\mathbf{b}, \sigma^2\mathbf{V})$ denotes a multivariate normal distribution with mean vector $\mathbf{b}$ and covariance matrix $\sigma^2\mathbf{V}$, and $IG(\alpha, \gamma)$ denotes an inverse gamma distribution that has a density function

$$p(u; \alpha, \gamma) = \frac{\gamma^\alpha}{\Gamma(\alpha)}\, u^{-\alpha-1} e^{-\gamma/u}, \quad u > 0, \quad \alpha > 0, \quad \gamma > 0. \tag{5.5}$$

In addition, we assume that $\{\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m\}$, $\{\boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta\}$, and σ_ϵ^2, three sets of parameters, are mutually independent. Furthermore, $\{\boldsymbol{\beta}_m, \sigma_m^2\}$ and $\{\boldsymbol{\phi}_m, \mathbf{P}_m\}$ are independent, and same are $\{\boldsymbol{\beta}_\delta, \sigma_\delta^2\}$ and $\{\boldsymbol{\phi}_\delta, \mathbf{P}_\delta\}$. As a result,

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}_m, \sigma_m^2) \cdot p(\boldsymbol{\phi}_m, \mathbf{P}_m) \cdot p(\boldsymbol{\beta}_\delta, \sigma_\delta^2) \cdot p(\boldsymbol{\phi}_\delta, \mathbf{P}_\delta) \cdot p(\sigma_\epsilon^2). \tag{5.6}$$

5.3.2 Posterior Distribution of Model Bias $\delta(\cdot)$

Given physical observations $\mathbf{y}^e$ and computer outputs $\mathbf{y}^m$, the posterior distribution of the model bias $\delta(D) = (\delta(\mathbf{x}_1), \cdots, \delta(\mathbf{x}_n))^T$ at any set $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ in the input space can be obtained by, according to Bayes' theorem,

$$p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m) = \int_{\boldsymbol{\theta}} p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\mathbf{y}^e, \mathbf{y}^m)\, \mathrm{d}\boldsymbol{\theta}. \tag{5.7}$$

We can easily derive the density $p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta})$.

Lemma 1 *Under the assumption that $D_e \subseteq D_m$, for any set D in the input space, the distribution of $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ is free of parameters $\boldsymbol{\beta}_m$, σ_m^2, $\boldsymbol{\phi}_m$, and $\mathbf{P}_m$.*

Proof: Assume that computer outputs are available at every point in D_e (i.e., $D_e \subseteq D_m$). Let $\mathbf{y}_{n_e}^m = (y^m(\mathbf{x}_1^e), \cdots, y^m(\mathbf{x}_{n_e}^e))^T$ be the vector of computer outputs at D_e. According to equation (5.1), we know that, given $\boldsymbol{\theta}$,

$$\begin{bmatrix} \delta(D) \\ \mathbf{y}^e \\ \mathbf{y}^m \end{bmatrix} \Bigg| \boldsymbol{\theta} \tag{5.8}$$

has a multivariate normal distribution with mean vector

$$\begin{bmatrix} \mathbf{F}_\delta(D)\boldsymbol{\beta}_\delta \\ \mathbf{F}_m(D_e)\boldsymbol{\beta}_m + \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta \\ \mathbf{F}_m(D_m)\boldsymbol{\beta}_m \end{bmatrix} \tag{5.9}$$

and covariance matrix

$$\begin{bmatrix} \sigma_\delta^2\mathbf{R}_\delta(D) & \sigma_\delta^2\mathbf{R}_\delta(D,D_e) & 0 \\ \sigma_\delta^2\mathbf{R}_\delta(D_e,D) & \sigma_m^2\mathbf{R}_m(D_e)+\sigma_\delta^2\mathbf{R}_\delta(D_e)+\sigma_\epsilon^2\mathbf{I}_{n_e} & \sigma_m^2\mathbf{R}_m(D_e,D_m) \\ 0 & \sigma_m^2\mathbf{R}_m(D_m,D_e) & \sigma_m^2\mathbf{R}_m(D_m) \end{bmatrix}. \tag{5.10}$$

Therefore,

$$\delta(D)|\mathbf{y}^e,\mathbf{y}^m,\boldsymbol{\theta} \sim N\Big(\mathbf{F}_\delta(D)\boldsymbol{\beta}_\delta + \mathbf{R}_\delta(D,D_e)(\mathbf{R}_\delta(D_e)+\tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e-\mathbf{y}^m_{n_e}-\mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta), \\ \sigma_\delta^2\left[\mathbf{R}_\delta(D)-\mathbf{R}_\delta(D,D_e)(\mathbf{R}_\delta(D_e)+\tau\mathbf{I}_{n_e})^{-1}\mathbf{R}_\delta(D_e,D)\right]\Big), \tag{5.11}$$

where $\tau=\sigma_\epsilon^2/\sigma_\delta^2$. As a result, the distribution of $\delta(D)|\mathbf{y}^e,\mathbf{y}^m,\boldsymbol{\theta}$ is free of parameters $\boldsymbol{\beta}_m$, σ_m^2, $\boldsymbol{\phi}_m$, and $\mathbf{P}_m$ (i.e., parameters in the Gaussian process $Y^m(\cdot)$). □

Theorem 2 *Under the assumptions that $D_e \subseteq D_m$ and $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ are known, for any set D in the input space,*

$$\delta(D)|\mathbf{y}^e,\mathbf{y}^m = \delta(D)|\mathbf{y}^e,\mathbf{y}^m_{n_e} \;\sim\; T_n(\nu_{\delta|e,m},\; \mu_{\delta|e,m}(D),\; \Sigma_{\delta|e,m}(D)), \tag{5.12}$$

where

$$\nu_{\delta|e,m} = n_e + 2\alpha_\delta, \tag{5.13a}$$

$$\mu_{\delta|e,m}(D) = \mathbf{H}_\delta^T\mathbf{A}_\delta\mathbf{v}_\delta + \mathbf{R}_\delta(D,D_e)(\mathbf{R}_\delta(D_e)+\tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e-\mathbf{y}^m_{n_e}), \tag{5.13b}$$

$$\Sigma_{\delta|e,m}(D) = \frac{Q_\delta^2}{\nu_{\delta|e,m}}\left[\mathbf{R}_\delta(D)-\mathbf{R}_\delta(D,D_e)(\mathbf{R}_\delta(D_e)+\tau\mathbf{I}_{n_e})^{-1}\mathbf{R}_\delta(D_e,D)+\mathbf{H}_\delta^T\mathbf{A}_\delta H_\delta\right], \tag{5.13c}$$

$$Q_\delta^2 = 2\gamma_\delta + (\mathbf{y}^e-\mathbf{y}^m_{n_e})^T(\mathbf{R}_\delta(D_e)+\tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e-\mathbf{y}^m_{n_e}) + \mathbf{b}_\delta\mathbf{V}_\delta^{-1}\mathbf{b}_\delta - \mathbf{v}_\delta^T\mathbf{A}_\delta\mathbf{v}_\delta, \tag{5.13d}$$

$$\mathbf{H}_\delta^T = \mathbf{F}_\delta(D) - \mathbf{R}_\delta(D,D_e)(\mathbf{R}_\delta(D_e)+\tau\mathbf{I}_{n_e})^{-1}\mathbf{F}_\delta(D_e), \tag{5.13e}$$

$$\mathbf{A}_\delta^{-1} = \mathbf{F}_\delta^T(D_e)(\mathbf{R}_\delta(D_e)+\tau\mathbf{I}_{n_e})^{-1}\mathbf{F}_\delta(D_e) + \mathbf{V}_\delta^{-1}, \tag{5.13f}$$

$$\mathbf{v}_\delta = \mathbf{F}_\delta^T(D_e)(\mathbf{R}_\delta(D_e)+\tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e-\mathbf{y}^m_{n_e}) + \mathbf{V}_\delta^{-1}\mathbf{b}_\delta. \tag{5.13g}$$

Proof: By Lemma 1, the distribution of $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ is free of parameters $\boldsymbol{\beta}_m$, σ_m^2, $\boldsymbol{\phi}_m$, and $\mathbf{P}_m$. Therefore, the posterior of $\delta(D)$ in equation (5.7) becomes

$$\begin{aligned} & p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m) \qquad (5.14) \\ =& \int_{\boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau} p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau)\, p(\boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau|\mathbf{y}^e, \mathbf{y}^m)\, \mathrm{d}\boldsymbol{\beta}_\delta\, \mathrm{d}\sigma_\delta^2\, \mathrm{d}\boldsymbol{\phi}_\delta\, \mathrm{d}\mathbf{P}_\delta\, \mathrm{d}\tau \end{aligned}$$

Assume that $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ are known. After integrating out $\boldsymbol{\beta}_\delta$ and σ_δ^2 (see Appendix 5.7.3), we have Theorem 2. □.

According to Theorem 2, under the assumptions that $D_e \subseteq D_m$ and parameters $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ are known, for any set D, the posterior of the model bias $\delta(D)$ is a multivariate noncentral t distribution with degree of freedom $\nu_{\delta|e,m}$, noncentrality parameter $\mu_{\delta|e,m}(D)$, and scale matrix $\Sigma_{\delta|e,m}(D)$. In other words, the posterior of the model bias $\delta(\cdot)$ is a noncentral t process. It is important to point out that the posterior of $\delta(\cdot)$ depends on only $\mathbf{y}^e$ and $\mathbf{y}_{n_e}^m$ (the vector of computer outputs at D_e) and computer outputs at $D_m - D_e$ do not help the prediction of $\delta(\cdot)$. The reason for such a result is because of two assumptions we made: (1) $D_e \subseteq D_m$ and (2) $Y^m(\cdot)$ and $\delta(\cdot)$ are mutually independent. In fact, the posterior distribution of $\delta(D)$ in equations (5.12) and (5.13) is the same as the one obtained by fitting a single Gaussian process with an additional experimental error to $\mathbf{y}^e - \mathbf{y}_{n_e}^m$, the difference of physical observations and computer outputs at D_e.

With Theorem 2, we can easily construct pointwise prediction intervals for the mode bias $\delta(\cdot)$ by using the fact that, at any $\mathbf{x}$,

$$\frac{\delta(\mathbf{x})|\mathbf{y}^e, \mathbf{y}^m - \mu_{\delta|e,m}(\mathbf{x})}{\sigma_{\delta|e,m}(\mathbf{x})} \;\sim\; T_1(\nu_{\delta|e,m}, 0, 1) \qquad (5.15)$$

where $T_1(\nu_{\delta|e,m}, 0, 1)$ is a univariate t distribution with degree of freedom $\nu_{\delta|e,m}$. This gives the $100(1-\alpha)\%$ prediction interval for the model bias $\delta(\mathbf{x})$ at any point $\mathbf{x}$ in the input space as

$$\mu_{\delta|e,m}(\mathbf{x}) \pm \sigma_{\delta|e,m}(\mathbf{x}) \cdot t_{\nu_{\delta|e,m}, \alpha/2} \qquad (5.16)$$

where $t_{\nu_{\delta|e,m},\alpha/2}$ is the upper $\alpha/2$ critical point of a univariate t distribution with degree of freedom $\nu_{\delta|e,m}$.

5.3.3 Posterior Distribution of Computer Output $Y^m(\cdot)$

The posterior distribution of the computer output $Y^m(D)$ can be derive in a very similar way in which the posterior distribution of $\delta(D)$ is derived. By Bayes' theorem, we have

$$p(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m) = \int_{\boldsymbol{\theta}} p(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\mathbf{y}^e, \mathbf{y}^m)\, \mathrm{d}\boldsymbol{\theta}. \tag{5.17}$$

Lemma 3 *Under the assumption that $D_e \subseteq D_m$, for any set D in the input space, the distribution of $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ is free of parameters $\boldsymbol{\beta}_\delta$, σ^2_δ, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and σ^2_ϵ.*

Proof: Assuming that $D_e \subseteq D_m$, in a similar way in which we derive the distribution of $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$, we can easily show that

$$\begin{aligned} Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta} \sim N\Big(&\mathbf{F}_m(D)\boldsymbol{\beta}_m + \mathbf{R}_m(D, D_m)\mathbf{R}_m^{-1}(D_m)(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m), \\ &\sigma^2_m \left[\mathbf{R}_m(D) - \mathbf{R}_m(D, D_m)\mathbf{R}_m^{-1}(D_m)\mathbf{R}_m(D_m, D)\right]\Big), \end{aligned} \tag{5.18}$$

which is free of parameters $\boldsymbol{\beta}_\delta$, σ^2_δ, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and σ^2_ϵ. □

Theorem 4 *Under the assumptions that $D_e \subseteq D_m$ and $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ are known, for any set D in the input space,*

$$Y^m(D)|\mathbf{y}^e, \mathbf{y}^m = Y^m(D)|\mathbf{y}^m \ \sim\ T_n(\nu_{m|m},\ \mu_{m|m}(D),\ \Sigma_{m|m}(D)), \tag{5.19}$$

where

$$\nu_{m|m} = n_m + 2\alpha_m, \tag{5.20a}$$

$$\mu_{m|m}(D) = \mathbf{H}_m^T \mathbf{A}_m \mathbf{v}_m + \mathbf{R}_m(D, D_m)\mathbf{R}_m^{-1}(D_m)\mathbf{y}^m, \tag{5.20b}$$

$$\Sigma_{m|m}(D) = \frac{Q_m^2}{\nu_{m|m}}\left[\mathbf{R}_m(D) - \mathbf{R}_m(D, D_m)\mathbf{R}_m^{-1}(D_m)\mathbf{R}_m(D_m, D) + \mathbf{H}_m^T \mathbf{A}_m H_m\right], \tag{5.20c}$$

$$Q_m^2 = 2\gamma_m + (\mathbf{y}^m)^T \mathbf{R}_m^{-1}(D_m)\mathbf{y}^m + \mathbf{b}_m \mathbf{V}_m^{-1}\mathbf{b}_m - \mathbf{v}_m^T \mathbf{A}_m \mathbf{v}_m, \tag{5.20d}$$

$$\mathbf{H}_m^T = \mathbf{F}_m(D) - \mathbf{R}_m(D, D_m)\mathbf{R}_m^{-1}(D_m)\mathbf{F}_m(D_m), \tag{5.20e}$$

$$\mathbf{A}_m^{-1} = \mathbf{F}_m^T(D_m)\mathbf{R}_m^{-1}(D_m)\mathbf{F}_m(D_m) + \mathbf{V}_m^{-1}, \tag{5.20f}$$

$$\mathbf{v}_m = \mathbf{F}_m^T(D_m)\mathbf{R}_m^{-1}(D_m)\mathbf{y}^m + \mathbf{V}_m^{-1}\mathbf{b}_m. \tag{5.20g}$$

Proof: By Lemma 3, the distribution of $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ is free of parameters $\boldsymbol{\beta}_\delta$, σ_δ^2, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and σ_ϵ^2. Therefore, equation (5.17) becomes

$$p(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m) \tag{5.21}$$
$$= \int_{\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m} p(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m)\, p(\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m|\mathbf{y}^e, \mathbf{y}^m)\, \mathrm{d}\boldsymbol{\beta}_m\, \mathrm{d}\sigma_m^2\, \mathrm{d}\boldsymbol{\phi}_m\, \mathrm{d}\mathbf{P}_m$$

Assume that $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ are known. After integrating out $\boldsymbol{\beta}_m$ and σ_m^2 (in a similar way we integrate out $\boldsymbol{\beta}_\delta$ and σ_δ^2), we have Theorem 4. □.

It is important to point out that $\mathbf{y}^e$ does not appear in the posterior distribution of $Y^m(D)$. In other words, the posterior distribution of $Y^m(D)$ is free of $\mathbf{y}^e$ and depends on only $\mathbf{y}^m$. We emphasize this fact by using a subscript $m|m$ instead of $m|e, m$. The reason for such a result is because of two assumptions we made: (1) $D_e \subseteq D_m$ and (2) $Y^m(\cdot)$ and $\delta(\cdot)$ are mutually independent. In fact, the posterior distribution of $Y^m(D)$ in equations (5.19) and (5.20) is the same as the one obtained by fitting a single Gaussian process to computer outputs $\mathbf{y}^m$ (Santner et al 2003).

Similar to the construction of pointwise prediction intervals for the model bias $\delta(\cdot)$, we can construct the $100(1-\alpha)\%$ prediction interval for the computer output $Y^m(\mathbf{x})$ at any point $\mathbf{x}$ in the input space as

$$\mu_{m|m}(\mathbf{x}) \pm \sigma_{m|m}(\mathbf{x}) \cdot t_{\nu_{m|m}, \alpha/2} \tag{5.22}$$

where $t_{\nu_{m|m},\alpha/2}$ is the upper $\alpha/2$ critical point of a univariate t distribution with degree of freedom $\nu_{m|m}$.

5.3.4 Posterior Distribution of Real System Output $Y^r(\cdot)$

We start the derivation of the posterior of the real system output $Y^r(\cdot)$ given physical observations $\mathbf{y}^e$ and computer outputs $\mathbf{y}^m$ by showing that the posteriors of the computer output $Y^m(\cdot)$ and the model bias $\delta(\cdot)$ are independent. We have shown that the posteriors of $\delta(\cdot)$ and $Y^m(\cdot)$ are both noncentral t processes. As a result, the posterior of $Y^r(\cdot)$ is the sum of two independent noncentral t processes.

Lemma 5 *Under the assumption that $D_e \subseteq D_m$, for any set D in the input space, $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ and $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ are independent.*

Proof: We know that, for any set D in the input space,

$$\begin{bmatrix} Y^r(D) \\ \mathbf{y}^e \\ \mathbf{y}^m \end{bmatrix} \Big| \boldsymbol{\theta} = \begin{bmatrix} Y^m(D) + \delta(D) \\ \mathbf{y}^e \\ \mathbf{y}^m \end{bmatrix} \Big| \boldsymbol{\theta} \tag{5.23}$$

has a multivariate normal distribution with mean vector

$$\begin{bmatrix} \mathbf{F}_m(D)\boldsymbol{\beta}_m + \mathbf{F}_\delta(D)\boldsymbol{\beta}_\delta \\ \mathbf{F}_m(D_e)\boldsymbol{\beta}_m + \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta \\ \mathbf{F}_m(D_m)\boldsymbol{\beta}_m \end{bmatrix} \tag{5.24}$$

and covariance matrix

$$\begin{bmatrix} \sigma_m^2\mathbf{R}_m(D) + \sigma_\delta^2\mathbf{R}_\delta(D) & \sigma_m^2\mathbf{R}_m(D, D_e) + \sigma_\delta^2\mathbf{R}_\delta(D, D_e) & \sigma_m^2\mathbf{R}_m(D, D_m) \\ \sigma_m^2\mathbf{R}_m(D_e, D) + \sigma_\delta^2\mathbf{R}_\delta(D_e, D) & \sigma_m^2\mathbf{R}_m(D_e) + \sigma_\delta^2\mathbf{R}_\delta(D_e) + \sigma_\epsilon^2\mathbf{I}_{n_e} & \sigma_m^2\mathbf{R}_m(D_e, D_m) \\ \sigma_m^2\mathbf{R}_m(D_m, D) & \sigma_m^2\mathbf{R}_m(D_m, D_e) & \sigma_m^2\mathbf{R}_m(D_m) \end{bmatrix}. \tag{5.25}$$

Therefore, under the assumption that $D_e \subseteq D_m$, $Y^r(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ has a multivariate normal distribution with mean vector

$$E[Y^r(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}] = \mathbf{F}_m(D)\boldsymbol{\beta}_m + \mathbf{R}_m(D, D_m)\mathbf{R}_m^{-1}(D_m)(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m) \quad (5.26)$$
$$+ \mathbf{F}_\delta(D)\boldsymbol{\beta}_\delta + \mathbf{R}_\delta(D, D_e)(\mathbf{R}_\delta(D_e) + \tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e - \mathbf{y}^m_{n_e} - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta),$$

and covariance matrix

$$\mathrm{Cov}(Y^r(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}) = \sigma_m^2 \left[\mathbf{R}_m(D) - \mathbf{R}_m(D, D_m)\mathbf{R}_m^{-1}(D_m)\mathbf{R}_m(D_m, D)\right] \quad (5.27)$$
$$+ \sigma_\delta^2 \left[\mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)(\mathbf{R}_\delta(D_e) + \tau\mathbf{I}_{n_e})^{-1}\mathbf{R}_\delta(D_e, D)\right].$$

Equation (5.27) together with equations (5.11) and (5.18) implies that, under the assumption that $D_e \subseteq D_m$,

$$\mathrm{Cov}(Y^r(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}) = \mathrm{Cov}(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}) + \mathrm{Cov}(\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}). \quad (5.28)$$

That is, $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ and $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}$ are uncorrelated. Therefore, they are independent. □

Lemma 6 *Under the assumption that $D_e \subseteq D_m$, for any set D in the input space, $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m$ and $\delta(D)|\mathbf{y}^e, \mathbf{y}^m$ are independent.*

Proof: The joint distribution of $Y^m(D)$ and $\delta(D)$ given $\mathbf{y}^e$ and $\mathbf{y}^m$ can be obtained by

$$\begin{aligned}
&p(Y^m(D), \delta(D)|\mathbf{y}^e, \mathbf{y}^m) \\
=&\int_{\boldsymbol{\theta}} p(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta})p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}^e, \mathbf{y}^m)\mathrm{d}\boldsymbol{\theta} \\
=&\int_{\boldsymbol{\theta}} p(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m)p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \sigma_\epsilon^2)p(\boldsymbol{\theta}|\mathbf{y}^e, \mathbf{y}^m)\mathrm{d}\boldsymbol{\theta} \\
=&\int_{\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m} p(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m)p(\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m|\mathbf{y}^e, \mathbf{y}^m)\mathrm{d}\boldsymbol{\beta}_m\mathrm{d}\sigma_m^2\mathrm{d}\boldsymbol{\phi}_m\mathrm{d}\mathbf{P}_m \\
&\cdot \int_{\boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau} p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau)p(\boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \sigma_\epsilon^2|\mathbf{y}^e, \mathbf{y}^m)\mathrm{d}\boldsymbol{\beta}_\delta\mathrm{d}\sigma_\delta^2\mathrm{d}\boldsymbol{\phi}_\delta\mathrm{d}\mathbf{P}_\delta\mathrm{d}\tau \\
=&p(Y^m(D)|\mathbf{y}^e, \mathbf{y}^m) \cdot p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m). \qquad (5.29)
\end{aligned}$$

That is, $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m$ and $\delta(D)|\mathbf{y}^e, \mathbf{y}^m$ are independent. The first step in equation (5.29) is due to Lemma 5, the second step is due to Lemmas 1 and 3, and the third step is due to the fact that

$$\begin{aligned} & p(\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau | \mathbf{y}^e, \mathbf{y}^m) \\ & \quad = p(\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m | \mathbf{y}^e, \mathbf{y}^m) \cdot p(\boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau | \mathbf{y}^e, \mathbf{y}^m) \end{aligned} \tag{5.30}$$

□

Theorem 7 *Under the assumptions that $D_e \subseteq D_m$ and $\boldsymbol{\phi}_m$, $\mathbf{P}_m$, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ are known, for any set D in the input space, the distribution of $Y^r(D)|\mathbf{y}^e, \mathbf{y}^m$ is the sum of two independent multivariate noncentral t distributions given in equations (5.12), (5.13), (5.19) and (5.20).*

Proof: By Theorem 2, $\delta(D)|\mathbf{y}^e, \mathbf{y}^m$ has a multivariate noncentral t distribution when $D_e \subseteq D_m$ and $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ are known. By Theorem 4, $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m$ has a multivariate noncentral t distribution when $D_e \subseteq D_m$ and $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ are known. Therefore, the posterior distribution of $Y^r(D)$ given $\mathbf{y}^e$ and $\mathbf{y}^m$ is the sum of two independent multivariate noncentral t distributions. □

The posterior mean and covariance matrix of $Y^r(D)$ can be easily obtained from equations (5.12), (5.13), (5.19) and (5.20)

$$E[Y^r(D)|\mathbf{y}^e, \mathbf{y}^m] = \mu_{r|e,m}(D) = \mu_{m|m}(D) + \mu_{\delta|e,m}(D), \tag{5.31a}$$

$$\mathrm{Cov}(Y^r(D)|\mathbf{y}^e, \mathbf{y}^m) = \Sigma_{r|e,m}(D) = \frac{\nu_{m|m}}{\nu_{m|m} - 2} \cdot \Sigma_{m|m}(D) + \frac{\nu_{\delta|e,m}}{\nu_{\delta|e,m} - 2} \cdot \Sigma_{\delta|e,m}(D). \tag{5.31b}$$

Pointwise prediction intervals for the real system output $Y^r(\cdot)$ can be constructed by using the fact that, for any $\mathbf{x}$,

$$\frac{Y^r(\mathbf{x})|\mathbf{y}^e, \mathbf{y}^m - \mu_{r|e,m}(\mathbf{x})}{\sigma_{r|e,m}(\mathbf{x})} \sim \frac{\sigma_{m|m}(\mathbf{x})}{\sigma_{r|e,m}(\mathbf{x})} T_1(\nu_{m|m}, 0, 1) + \frac{\sigma_{\delta|e,m}(\mathbf{x})}{\sigma_{r|e,m}(\mathbf{x})} T_1(\nu_{\delta|e,m}, 0, 1). \tag{5.32}$$

This gives the $100(1-\alpha)\%$ prediction interval for the real system output $Y^r(\mathbf{x})$ at any point $\mathbf{x}$ in the input space as

$$\mu_{r|e,m}(\mathbf{x}) \pm c_{\alpha/2} \cdot \sigma_{r|e,m}(\mathbf{x}) \tag{5.33}$$

where the critical point $c_{\alpha/2}$ can be estimated from observations randomly generated from the two independent univariate t distributions, $T_1(\nu_{\delta|e,m}, 0, 1)$ and $T_1(\nu_{m|m}, 0, 1)$.

5.4 When $D_e \not\subseteq D_m$ and $\boldsymbol{\phi}_m$, $\mathbf{P}_m$, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ are unknown

When deriving the posteriors of the model bias $\delta(\cdot)$ and the real system output $Y^r(\cdot)$, we assume that: (1) computer outputs are available at D_e (i.e., $D_e \subseteq D_m$), and (2) parameters $\boldsymbol{\phi}_m$, $\mathbf{P}_m$, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ are known. These two assumptions are not always true especially the second one. Below we describe how we handle those two situations.

5.4.1 Prediction of $Y^m(D_e - D_m)$

When $D_e \not\subseteq D_m$, we could either simulate computer outputs at $D_e - D_m$ (the set of points in D_e but not in D_m) by running the computer model if it is not too expensive to do so or we could predict computer outputs at $D_e - D_m$ using the posterior mean of $Y^m(D_e - D_m)$

$$\hat{Y}^m(D_e - D_m) = \mu_{m|m}(D_e - D_m). \tag{5.34}$$

We often have a computer design set D_m large enough such that the prediction $\hat{Y}^m(D_e - D_m)$ is quite accurate and there is little loss in using $\hat{Y}^m(D_e - D_m)$ as real computer outputs $Y^m(D_e - D_m)$. Having computer outputs at all points in D_e, we can use the results derived in previous section.

5.4.2 Estimation of $\boldsymbol{\phi}_m$, $\mathbf{P}_m$, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ

The posterior of the model bias $\delta(D)$ in equation (5.12) is in fact the conditional posterior of $\delta(D)$ given $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ. To get the marginal posterior of $\delta(D)$, we

have to integrate out $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ, which often is difficult to derive analytically or computationally prohibitive. Therefore, instead of deriving the marginal posterior of $\delta(D)$ by integrating out $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ, we estimate $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ and then treat the estimates as if they are the true values of $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ. Same is done for $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ in the posterior of the computer output $Y^m(D)$. Instead of integrating out $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$, we estimate $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ and treat the estimates as if they are the true values of $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$.

• Maximum Likelihood (ML) Estimates

The ML estimates of $\boldsymbol{\phi}_m$, $\mathbf{P}_m$, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ maximize the likelihood

$$p(\mathbf{y}^e, \mathbf{y}^m | \boldsymbol{\phi}_m, \mathbf{P}_m, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau). \tag{5.35}$$

Since

$$p(\mathbf{y}^e, \mathbf{y}^m | \boldsymbol{\phi}_m, \mathbf{P}_m, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau) = p(\mathbf{y}^e | \mathbf{y}^m, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \tau) \cdot p(\mathbf{y}^m | \boldsymbol{\phi}_m, \mathbf{P}_m), \tag{5.36}$$

and (see Appendix 5.7.4)

$$\begin{aligned} p(\mathbf{y}^m | \boldsymbol{\phi}_m, \mathbf{P}_m) &\propto |\mathbf{R}_m(D_m)|^{-\frac{1}{2}} \cdot |\mathbf{A}_m|^{\frac{1}{2}} \qquad (5.37a) \\ &\cdot \left[2\gamma_m + (\mathbf{y}^m)^T \mathbf{R}_m^{-1}(D_m)\mathbf{y}^m + \mathbf{b}_m^T \mathbf{V}_m^{-1} \mathbf{b}_m - \mathbf{v}_m^T \mathbf{A}_m \mathbf{v}_m\right]^{-\frac{n_m}{2} - \alpha_m}, \\ p(\mathbf{y}^e | \mathbf{y}^m, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \sigma_\epsilon^2) &\propto |\boldsymbol{\phi}_\delta + \tau \mathbf{I}_{n_e}|^{-\frac{1}{2}} \cdot |\mathbf{A}_\delta|^{\frac{1}{2}} \qquad (5.37b) \\ &\cdot \left[2\gamma_\delta + (\mathbf{y}^e - \mathbf{y}_{n_e}^m)^T (\mathbf{R}_\delta(D_e) + \tau \mathbf{I}_{n_e})^{-1} (\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{b}_\delta^T \mathbf{V}_\delta^{-1} \mathbf{b}_\delta - \mathbf{v}_\delta^T \mathbf{A}_\delta \mathbf{v}_\delta\right]^{-\frac{n_e}{2} - \alpha_\delta}, \end{aligned}$$

we estimate $\boldsymbol{\phi}_m$, $\mathbf{P}_m$, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ in two steps:

– First, estimate $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ by minimizing

$$\begin{aligned} &\log(|\mathbf{R}_m(D_m)|) - \log(|\mathbf{A}_m|) \qquad (5.38) \\ &+(n_m + 2\alpha_m) \cdot \log\left[2\gamma_m + (\mathbf{y}^m)^T \mathbf{R}_m^{-1}(D_m)\mathbf{y}^m + \mathbf{b}_m^T \mathbf{V}_m^{-1} \mathbf{b}_m - \mathbf{v}_m^T \mathbf{A}_m \mathbf{v}_m\right], \end{aligned}$$

which uses only computer outputs $\mathbf{y}^m$.

– Next, estimate $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ by minimizing

$$\log(|\mathbf{R}_\delta(D_e) + \tau\mathbf{I}_{n_e}|) - \log(|\mathbf{A}_\delta|) + (n_e + 2\alpha_\delta) \tag{5.39}$$
$$\cdot \log\left[2\gamma_\delta + (\mathbf{y}^e - \mathbf{y}^m_{n_e})^T(\mathbf{R}_\delta(D_e) + \tau\mathbf{I}_{n_e})^{-1}(\mathbf{y}^e - \mathbf{y}^m_{n_e}) + \mathbf{b}_\delta^T\mathbf{V}_\delta^{-1}\mathbf{b}_\delta - \mathbf{v}_\delta^T\mathbf{A}_\delta\mathbf{v}_\delta\right],$$

which uses both physical observations $\mathbf{y}^e$ and computer outputs $\mathbf{y}^m$. When computer outputs at some points in D_e are not available, one can use the posterior means of $Y^m(\cdot)$ at those points instead.

• Markov Chain Monte Carlo (MCMC) Estimates

A common problem with maximum likelihood estimation is that the minimization of either (5.38) or (5.39) is highly unstable, unable to converge or converging to a local minimum. An alternative is to use Markov Chain Monte Carlo (MCMC) algorithms (Lange 1999, Liu 2001) to estimate the posterior distributions of $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$, and then use the posterior means or modes of $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ as the estimates of $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ (Bayarri et al 2002, Qian and Wu 2005). Same can be done for $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ.

• Minimum Mean Squared Prediction Error (MMSPE) Estimates

Another way to estimate $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ is to find the values of $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ that minimize the mean squared prediction errors on a hold-out sample (i.e., a sample that is not used for the calculation of the posterior of $Y^m(\cdot)$). The resulting estimates are the MMSPE estimates of $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ (Santner et al 2003). When there are no enough computer outputs to afford a hold-out sample, cross validation can be used instead. Same can be done for $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ.

5.5 A Bayesian Validation Procedure

In this section, we present a complete Bayesian validation procedure based on the results derived above. This procedure consists of the following steps:

(1). Collect data.

Both physical observations and computer outputs are essential to model validation. We should collect as many as possible physical observations at input points as close to the input region of interest as possible. Compared to physical observations, computer outputs are less costly and should be computed at points where physical observations are available and close to if not within the input region of interest. After data collection is done, we should have D_e, $\mathbf{y}^e$, D_m, and $\mathbf{y}^m$.

(2). Determine the priors of model parameters.

One advantage the Bayesian approach has is its ability to take into account scientific knowledge and past information on the real system in the form of prior distributions for unknown model parameters. For the model in equation (5.1), a priori knowledge can be expressed as priors for parameters in $\boldsymbol{\theta}$. When there is little a priori knowledge, we use the "vague" priors:

$$\sigma_m^2 \sim IG(2,1),\ \boldsymbol{\beta}_m|\sigma_m^2 \sim N(\mathbf{0}_{q_m}, \sigma_m^2\mathbf{I}_{q_m}),\ \sigma_\delta^2 \sim IG(2,1),\ \boldsymbol{\beta}_\delta|\sigma_\delta^2 \sim N(\mathbf{0}_{q_\delta}, \sigma_\delta^2\mathbf{I}_{q_\delta}), \tag{5.40}$$

where $\mathbf{0}_k$ is a $k \times 1$ vector of zeros, and $\mathbf{I}_k$ is a $k \times k$ identity matrix.

(3). Estimate $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ and calculate the posterior of $Y^m(D)$.

As mentioned before, the posterior of $Y^m(D)$ in equations (5.19) and (5.20) is conditional on $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$. To get the marginal posterior of $Y^m(D)$, we need to integrate out $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$, which is often difficult or impossible to do. As a result, we estimate $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$ and treat the estimates as their true values. There are many ways to estimate $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$. Three methods, ML, MCMC, and MMSPE, are mentioned in section 5.4.2. With the estimates of $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$, the posterior of $Y^m(D)$ is given by equations (5.12) and (5.13).

(4). Estimate $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ.

Same as for $\boldsymbol{\phi}_m$ and $\mathbf{P}_m$, instead of integrating out $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ to give the marginal posterior of $\delta(D)$, we estimate $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ and treat the estimates as their true values.

(5). Calculate the posterior of $\delta(D)$.

Having physical observations $\mathbf{y}^e$, computer outputs $\mathbf{y}^m$ (computed from the computer model or predicted using the posterior distribution of $Y^m(\cdot)$), the priors for parameters, and the estimates of $\boldsymbol{\phi}_m$, $\mathbf{P}_m$, $\boldsymbol{\phi}_\delta$, $\mathbf{P}_\delta$, and τ, we can calculate the posterior distribution of $\delta(D)$ using equations (5.12) and (5.13).

(6). Calculate the posterior of $Y^r(D)$.

As proven in section 5.3.4, under the assumption that $D_e \subseteq D_m$, the posteriors of $Y^m(D)$ and $\delta(D)$ are independent multivariate noncentral t distributions. As a result, the posterior of $Y^r(D)$ is the sum of two independent multivariate noncentral t distributions. The posterior mean and covariance matrix of $Y^r(D)$ are given in equation (5.31).

(7). Validate the computer model.

Having the posterior distributions of $\delta(\cdot)$ and $Y^r(\cdot)$, we can construct prediction intervals for $\delta(\mathbf{x})$ and $Y^r(\mathbf{x})$ at any $\mathbf{x}$ (see equations (5.16) and (5.33)). Both prediction intervals can be used for the validation of the computer model. The traditional approach is to reject the computer model at $\mathbf{x}$ if the prediction interval for $\delta(\mathbf{x})$ (or $Y^r(\mathbf{x})$) does not contain zero (or $Y^m(\mathbf{x})$) and fail to reject otherwise. There are two problems with this approach. First, we tend to reject the computer model at points where more physical observations are available and therefore prediction intervals are narrower while fail to reject the computer model at points where fewer or no physical observations are available and therefore prediction intervals are wider. Second, failing to reject the computer model

at a point does not necessarily mean that the computer model is acceptable. To overcome those two problems, we propose a new way for model validation based on prediction intervals.

Let $l_r(\mathbf{x})$ and $u_r(\mathbf{x})$ be the lower and upper bounds of the $100(1-\alpha)\%$ prediction interval for $Y^r(\mathbf{x})$ respectively. According to equation (5.33), we have

$$l_r(\mathbf{x}) = \mu_{r|e,m}(\mathbf{x}) - c_{\alpha/2} \cdot \sigma_{r|e,m}(\mathbf{x}), \tag{5.41a}$$

$$u_r(\mathbf{x}) = \mu_{r|e,m}(\mathbf{x}) + c_{\alpha/2} \cdot \sigma_{r|e,m}(\mathbf{x}). \tag{5.41b}$$

Define

$$\Delta_{min}(\mathbf{x}) = \begin{cases} 0, & \text{if } Y^m(\mathbf{x}) \in (l_r(\mathbf{x}),\, u_r(\mathbf{x})), \\ \min\left\{|Y^m(\mathbf{x}) - l_r(\mathbf{x})|, |Y^m(\mathbf{x}) - u_r(\mathbf{x})|\right\}, & \text{otherwise.} \end{cases} \tag{5.42}$$

and

$$\Delta_{max}(\mathbf{x}) = \max\left\{|Y^m(\mathbf{x}) - l_r(\mathbf{x})|, |Y^m(\mathbf{x}) - u_r(\mathbf{x})|\right\}. \tag{5.43}$$

In other words, $\Delta_{min}(\mathbf{x})$ and $\Delta_{max}(\mathbf{x})$ are the minimum and maximum possible deviations of $Y^m(\mathbf{x})$ from the real system output $Y^r(\mathbf{x})$. Let Δ_0 be a pre-specified threshold.

- If $\Delta_{min}(\mathbf{x}) > \Delta_0$, we reject the computer model as an acceptable representative of the real system at $\mathbf{x}$.
- If $\Delta_{max}(\mathbf{x}) < \Delta_0$, we conclude that the computer model is acceptable at $\mathbf{x}$.
- If $\Delta_{min}(\mathbf{x}) \leq \Delta_0 \leq \Delta_{max}(\mathbf{x})$, no conclusion can be reached and more physical observations are needed.

The validation based on the prediction interval of $\delta(\mathbf{x})$ is similar. Let $l_\delta(\mathbf{x})$ and $u_\delta(\mathbf{x})$ be the lower and upper bounds of the $100(1-\alpha)\%$ prediction interval for

$\delta(\mathbf{x})$ respectively. According to equation (5.16), we have

$$l_\delta(\mathbf{x}) = \mu_{\delta|e,m}(\mathbf{x}) - \sigma_{\delta|e,m}(\mathbf{x}) \cdot t_{\nu_{\delta|e,m},\alpha/2}, \tag{5.44a}$$

$$u_\delta(\mathbf{x}) = \mu_{\delta|e,m}(\mathbf{x}) + \sigma_{\delta|e,m}(\mathbf{x}) \cdot t_{\nu_{\delta|e,m},\alpha/2}. \tag{5.44b}$$

The formulas for $\Delta_{min}(\mathbf{x})$ and $\Delta_{max}(\mathbf{x})$ are now given as

$$\Delta_{min}(\mathbf{x}) = \begin{cases} 0, & \text{if } 0 \in (l_\delta(\mathbf{x}),\, u_\delta(\mathbf{x})), \\ \min\left\{|l_\delta(\mathbf{x})|, |u_\delta(\mathbf{x})|\right\}, & \text{otherwise.} \end{cases} \tag{5.45}$$

and

$$\Delta_{max}(\mathbf{x}) = \max\left\{|l_\delta(\mathbf{x})|, |u_\delta(\mathbf{x})|\right\}. \tag{5.46}$$

Having $\Delta_{min}(\mathbf{x})$ and $\Delta_{max}(\mathbf{x})$, with a pre-specified Δ_0, we can make decisions about the computer model in the same way as before.

The $\Delta_{min}(\mathbf{x})$ and $\Delta_{max}(\mathbf{x})$ based on $l_\delta(\mathbf{x})$ and $u_\delta(\mathbf{x})$ will be exactly the same as the $\Delta_{min}(\mathbf{x})$ and $\Delta_{max}(\mathbf{x})$ based on $l_r(\mathbf{x})$ and $u_r(\mathbf{x})$ if $\mathbf{x} \in D_m$ and sightly different from the latter if $\mathbf{x} \notin D_m$. The reason is because that $l_r(\mathbf{x})$ and $u_r(\mathbf{x})$ consider the variation in the computer output prediction while $l_\delta(\mathbf{x})$ and $u_\delta(\mathbf{x})$ do not. However, this difference is often very small and negligible since enough computer outputs are often available to guarantee a small prediction variation.

5.6 Numerical Experiments

We use three examples to illustrate the proposed Bayesian approach. For the first two examples, $D_e = D_m$. That is, the design set for the physical experiment is exactly the same as that for the computer experiment. For the third example, $D_e \cap D_m = \emptyset$. That is, the two design sets have no common points, which means that we need to predict computer outputs at D_e before computing the posterior distribution of $\delta(\mathbf{x})$. For all three examples, we compare the results from the proposed Bayesian approach to those from an approach proposed by Kennedy and O'Hagan (2001). Kennedy and

O'Hagan (2001) also use an exponential correlation function as shown in equations (5.2) and (5.3) but have the values of $P_k, k = 1, \cdots, p$ fixed at 2. In order for the results from two approaches to be comparable, we run our Bayesian approach with $P_{m,k} = P_{\delta,k} = 2, k = 1, \cdots, p$.

5.6.1 Example 1: Fluidized-Bed Coating

Dewettinck et al. (1999) described a Glatt GPCG-1 fluidized-bed unit for coating food products and several computer models developed for this unit to calculate the steady-state thermodynamic operation point. They also reported 28 physical observations of the steady-state outlet air temperature under different values of six factors: the room humidity (H_r) and temperature (T_r), the inlet air temperature (T_a), the flow rate of the coating liquid (R_f), the pressure of the atomization air (P_a), and the fluid velocity of the fluidization air (V_f). The 28 physical observations along with the corresponding values of the six factors are displayed in Table 5.1, which also contains the steady-state outlet air temperatures computed from one computer model.

As the six factors have different units, we normalize each factor before applying the proposed Bayesian approach. We then divide the data into two parts, reserving runs 4, 15, 17, 21, 23, 25, 26, and 28 as the testing data (the same partition used by Qian and Wu (2005)). We assume a constant mean for both Gaussian processes, $Y^m(\cdot)$ and $\delta(\cdot)$ and use the priors in equation (5.40). After getting the ML estimates of $\boldsymbol{\phi}_\delta$ and τ, we compute the predictions of $\delta(\cdot)$ and $Y^r(\cdot)$ for the runs in the testing data. The results are displayed in Tables 5.2 and 5.3, which also contain results from another two approaches:

- The proposed Bayesian approach with $Y^m(\cdot) \equiv 0$. The purpose is to see if including computer outputs improves the predictions of $Y^r(\cdot)$.
- The approach by Kennedy and O'Hagan (2001)

The δ column in Table 5.2 contains observed model biases that equal to the differences

of physical observations and computer outputs. The zeros in the $\hat{\delta}_m$ column are obtained when we assume that there are no biases in computer outputs. The last three columns in Table 5.2 give the model biases computed using the three methods mentioned above respectively. The root mean squared prediction error (RMSPE) is used to compare the performances of the three approaches. Both Tables 5.2 and 5.3 show that:

- Treating the computer model as it has no bias leads to a large RMSPE (1.9812).
- Using only physical observations leads to a larger RMSPE (1.5663) than using both physical observations and computer outputs (0.6856).
- The proposed method using both computer outputs and physical observations yields a slightly smaller RMSPE (0.6856) than Kennedy and O'Hagan's method (0.7119).

5.6.2 Example 2: Linear Cellular Alloys

This example is taken from Qian and Wu (2005). The quantity of interest is the steady-state heat transfer rate of a heat exchanger used in an electrical cooling application. Factors considered include the mass flow rate (m) and temperature (T_{in}) of the inlet air, the temperature of the heat source (T_{wall}), and the thermal conductivity (k). The data are presented in Table 5.4, in which the y^e and y^m columns are outputs of two computer models. The computer model generating y^e is more accurate and complicated than the one generating y^m. We use the notation y^e just for consistency. The y^e's here can be considered as physical observations without measurement errors (i.e., $\tau = 0$).

Same as in Example 1, we first normalize the factors, and then divide the data into two parts, using 24 runs as the training data and 8 runs (runs 1, 4, 9, 11, 13, 23, 25, and 27) as the testing data (the same partition used by Qian and Wu (2005)). Using

Table 5.1: The Fluidized-Bed Coating Example

run	$H_r(\%)$	$T_r(^oC)$	$T_a(^oC)$	$R_f(g/min)$	P_a(bar)	$V_f(m/s)$	y^e	y^m
1	51.0	20.7	50	5.52	2.5	3.0	30.4	31.5
2	46.4	21.3	60	5.53	2.5	3.0	37.6	38.5
3	46.6	19.2	70	5.53	2.5	3.0	45.1	45.5
4	53.1	21.1	80	5.51	2.5	3.0	50.2	52.6
5	52.0	20.4	90	5.21	2.5	3.0	57.9	59.9
6	45.6	21.4	60	7.25	2.5	3.0	32.9	34.6
7	47.3	19.5	70	7.23	2.5	3.0	39.5	41.0
8	53.3	21.4	80	7.23	2.5	3.0	45.6	48.5
9	44.0	20.1	70	8.93	2.5	3.0	34.2	36.6
10	52.3	21.6	80	8.91	2.5	3.0	41.1	44.3
11	55.0	20.2	80	7.57	1.0	3.0	45.7	49.0
12	54.0	20.6	80	7.58	1.5	3.0	44.6	48.4
13	50.8	21.1	80	7.40	2.0	3.0	44.7	48.4
14	48.0	21.2	80	7.43	2.5	3.0	44.0	48.0
15	42.8	22.4	80	7.51	3.0	3.0	43.3	47.5
16	55.7	20.8	50	3.17	1.0	3.0	37.0	38.0
17	55.2	20.7	50	3.18	1.5	3.0	37.2	38.5
18	54.4	20.7	50	3.19	2.0	3.0	37.1	37.5
19	55.4	19.8	50	3.20	2.5	3.0	36.9	38.5
20	52.9	20.0	50	3.19	3.0	3.0	36.8	37.2
21	28.5	18.3	80	7.66	2.5	3.0	46.0	47.3
22	26.1	19.0	80	7.69	2.5	4.0	54.7	56.2
23	24.2	18.9	80	7.69	2.5	4.5	57.0	58.7
24	25.4	18.5	80	7.70	2.5	5.0	58.9	60.5
25	45.1	19.6	50	3.20	2.5	3.0	35.9	37.1
26	43.1	20.3	50	3.23	2.5	4.0	40.3	40.8
27	42.7	20.4	50	3.20	2.5	4.5	41.9	42.3
28	38.7	21.6	50	3.22	2.5	5.0	43.1	43.3

Table 5.2: The Fluidized-Bed Coating Example: prediction of $\delta(\mathbf{x})$

run	y^e	y^m	δ $(y^e - y^m)$	$\hat{\delta}_m$ $(y^m - y^m)$	$\hat{\delta}_e$ $(\hat{y}_e^r - y^m)$	$\hat{\delta}$ (Proposed Method)	$\hat{\delta}_{KO}$ $(\hat{y}_{KO}^r - y^m)$
4	50.20	52.60	-2.4000	0	-2.2927	-2.2108	-2.3509
15	43.30	47.50	-4.2000	0	-3.0360	-2.3824	-2.4837
17	37.20	38.50	-1.3000	0	-1.4962	-0.7623	-0.8576
21	46.00	47.30	-1.3000	0	2.6269	-1.3062	-1.2199
23	57.00	58.70	-1.7000	0	-1.6524	-1.5950	-1.5707
25	35.90	37.10	-1.2000	0	0.4054	-1.0197	-0.5741
26	40.30	40.80	-0.5000	0	-0.2511	-0.6935	-0.3865
28	43.10	43.30	-0.2000	0	0.1998	-0.4243	-0.8945
			RMSPE	1.9812	1.5663	0.6856	0.7119

Table 5.3: The Fluidized-Bed Coating Example: prediction of $Y^r(\mathbf{x})$

run	y^e	y^m	$\hat{y}_e^r$	$\hat{y}^r$ (Proposed Method)	$\hat{y}_{KO}^r$
4	50.20	52.60	50.3073	50.3892	50.2491
15	43.30	47.50	44.4640	45.1176	45.0163
17	37.20	38.50	37.0038	37.7377	37.6424
21	46.00	47.30	49.9269	45.9938	46.0801
23	57.00	58.70	57.0476	57.1050	57.1293
25	35.90	37.10	37.5054	36.0803	36.5259
26	40.30	40.80	40.5489	40.1065	40.4135
28	43.10	43.30	43.4998	42.8757	42.4055
	RMSPE		1.9812	1.5663	0.6856	0.7119

the priors in equation (5.40) and assuming a linear mean for both $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$, we predict $\delta(\cdot)$ and $Y^r(\cdot)$ for the runs in the testing data using the three methods

- The proposed Bayesian approach
- The proposed Bayesian approach with $Y^m(\cdot) \equiv 0$
- The approach by Kennedy and O'Hagan (2001)

The results are displayed in Tables 5.5 and 5.6. Both show that using the outputs from both computer models leads to a significant decrease in the value of RMSPE (from 6.3059 to 2.4818). In addition, the proposed Bayesian approach yields a slightly smaller RMSPE (2.4818) than the approach by Kennedy and O'Hagan (2.8678).

5.6.3 Example 3: Compressible Shear Layer

Oberkampf and Barone (2004) gave a detailed description on free shear layers. Basically, a free shear layer is formed when two streams with distinct velocities and temperatures separated by a thin splitter plate mix downstream of the splitter plate trailing edge. Of particular interest is the behavior of the free shear layer as the convective Mach number M_c (representing the mixing of the two streams) increases for fixed velocity and temperature ratios. The behavior of the shear layer is usually represented by the compressibility factor defined as the ratio of the compressible growth rate (d_c) to the incompressible growth rate (d_i) at the same velocity and temperature ratios. Oberkampf and Barone (2004) reported both physical observations and computer outputs on the compressibility factor of shear layers. The physical observations are from several independent sources as shown by the column `Source` in Table 5.7. The eleven computer outputs in Table 5.8 are the values of the compressibility factor computed as M_c changes from 0.1 to 1.5 with an increase of 0.14 every time. The interval $(0, 1.5]$ is chosen to span the M_c range of the physical data.

Since little is known about model parameters beforehand, we use the priors in equation (5.40). As computer outputs at D_e are unavailable, we first compute the

Table 5.4: The Linear Cellular Alloys Example

run	$m(kg/s)$	$T_{in}(K)$	$k(W/mk)$	$T_{wall}(K)$	y^e	y^m
1	0.000500	293.15	362.7	393.1	23.5	25.6
2	0.000550	315.00	310.0	365.0	20.1	21.2
3	0.000552	293.53	318.6	388.3	10.2	11.4
4	0.000557	290.18	298.3	377.5	15.3	15.0
5	0.000560	277.01	355.0	374.0	18.4	18.6
6	0.000566	285.77	266.7	367.3	20.5	20.7
7	0.000578	302.17	358.1	343.7	30.1	30.2
8	0.000580	272.26	211.7	333.6	18.2	18.1
9	0.000589	278.16	225.8	351.8	24.7	25.0
10	0.000594	279.54	258.5	360.1	19.1	17.9
11	0.000603	296.75	323.1	399.4	25.0	24.2
12	0.000612	280.83	291.5	394.7	16.9	17.5
13	0.000615	300.28	270.7	335.8	22.3	22.5
14	0.000620	275.00	225.0	340.0	19.6	25.1
15	0.000626	284.89	350.5	352.3	23.3	18.9
16	0.000627	287.60	244.0	382.5	14.4	18.2
17	0.000652	298.04	304.0	361.6	21.3	13.8
18	0.000657	294.24	330.6	375.5	36.1	29.1
19	0.000670	303.07	321.4	370.5	25.4	22.2
20	0.000680	313.28	259.1	350.0	22.9	21.6
21	0.000683	287.05	227.3	358.2	34.5	30.9
22	0.000689	272.70	260.9	355.4	14.8	13.1
23	0.000694	278.35	212.8	376.2	18.8	16.4
24	0.000698	277.52	299.4	338.4	32.9	31.1
25	0.000700	288.15	300.0	400.0	17.4	13.5
26	0.000711	292.26	273.3	392.5	7.5	7.0
27	0.000714	283.08	306.7	344.3	42.9	35.5
28	0.000730	285.51	217.7	383.9	22.0	20.9
29	0.000738	295.01	295.0	347.2	19.8	25.5
30	0.000741	270.95	275.2	356.9	4.5	10.2
31	0.000751	287.99	326.0	354.1	47.0	36.6
32	0.000757	300.64	235.0	391.7	25.8	27.2

Table 5.5: The Linear Cellular Alloys Example: prediction of $\delta(\mathbf{x})$

run	y^e	y^m	δ $(y^e - y^m)$	$\hat{\delta}_m$ $(y^m - y^m)$	$\hat{\delta}_e$ $(\hat{y}_e^r - y^m)$	$\hat{\delta}$ (Proposed Method)	$\hat{\delta}_{KO}$ $(\hat{y}_{KO}^r - y^m)$
1	23.54	25.61	-2.0700	0	-8.7159	0.9291	2.9470
4	15.29	15.03	0.2600	0	3.6921	-1.1039	-0.3477
9	24.68	25.02	-0.3400	0	-7.6888	0.0524	-1.3082
11	24.96	24.20	0.7600	0	-7.9413	1.1079	2.5524
13	22.30	22.48	-0.1800	0	2.9055	-1.7799	-0.1964
23	18.78	16.40	2.3800	0	-1.2296	0.6742	-1.6294
25	17.41	13.54	3.8700	0	2.1570	1.1581	1.1535
27	42.93	35.53	7.4000	0	-2.9424	2.3681	3.8434
			RMSE	3.1717	6.3059	2.4818	2.8678

Table 5.6: The Linear Cellular Alloys Example: prediction of $Y^r(\mathbf{x})$

run	y^e	y^m	$\hat{y}_e^r$	$\hat{y}^r$ (Proposed Method)	$\hat{y}_{KO}^r$
1	23.54	25.61	16.8941	26.5391	28.5570
4	15.29	15.03	18.7221	13.9261	14.6823
9	24.68	25.02	17.3312	25.0724	23.7118
11	24.96	24.20	16.2587	25.3079	26.7524
13	22.30	22.48	25.3855	20.7001	22.2836
23	18.78	16.40	15.1704	17.0742	14.7706
25	17.41	13.54	15.6970	14.6981	14.6935
27	42.93	35.53	32.5876	37.8981	39.3734
	RMSE	3.1717	6.3059	2.4818	2.8678

posterior of $Y^m(x)$, $p(Y^m(x)|\mathbf{y}^m, \phi_m)$, which requires the estimate of ϕ_m. Notice that we use ϕ_m and x instead of $\boldsymbol{\phi}_m$ and $\mathbf{x}$ in boldface since this example has only one input variable. The MMSPE estimate of ϕ_m is used for this example. That is, the value of ϕ_m is chosen to minimize

$$\text{RMSPE} = \sqrt{\frac{1}{n_m}\sum_{i=1}^{n_m}[y^m(x_i') - \hat{y}^m(x_i')]^2}, \tag{5.47}$$

where $x_i',\ i = 1, \cdots, n_m$ are points in D_m, $\hat{y}^m(x_i') = E[Y^m(x_i')|\mathbf{y}^m_{-i}, \phi_m]$ is a function of ϕ_m (see equation (5.20b)), and $\mathbf{y}^m_{-i}$ contains all available computer outputs except $y^m(x_i')$. In other words, $\hat{y}^m(x_i')$'s are the leave-one-out cross validation predictions of computer outputs at D_m. Figure 5.1 displays the RMSPE as a function of ϕ_m and contains two curves, one for $f_m(x) \equiv 1$ (i.e., the mean of the Gaussian process $Y^m(x)$ is constant) and the other for $\mathbf{f}_m(x) = (1, x)^T$ (i.e., the mean of the Gaussian process $Y^m(x)$ is a linear function of x). Figure 5.1 shows that assuming a linear mean for $Y^m(x)$ leads to a smaller RMSPE, and the minimum RMSPE is achieved at $\hat{\phi}_m = 2$. Treating the MMSPE estimate of ϕ_m as its true value, we compute the posterior of $Y^m(x)$ using equations (5.19) and (5.20) and display the posterior mean of $Y^m(x)$ along with the corresponding 95% prediction interval in Figure 5.2, which shows a rather small posterior variance except at the two ends of the M_c range. Figure 5.2 also displays physical observations as circles.

The next step is to find the MMSPE estimates of ϕ_δ and τ together by minimizing

$$\text{RMSPE} = \sqrt{\frac{1}{n_e}\sum_{i=1}^{n_e}\left[y^e(x_i) - \hat{y}^m(x_i) - \hat{\delta}(x_i)\right]^2}, \tag{5.48}$$

where $x_i,\ i = 1, \cdots, n_e$ are points in D_e, $\hat{y}^m(x_i) = E[Y^m(x_i)|\mathbf{y}^m, \hat{\phi}_m] = \mu_{m|m}(x_i)$, and $\hat{\delta}(x_i)$ is the ten-fold cross validation prediction of $\delta(x_i)$ and is a function of ϕ_δ and τ. The results show that the RMSPE reaches minimum at $\hat{\phi}_\delta = 1.4$ and $\hat{\tau} = 0.02$ with the mean of the Gaussian process $\delta(x)$ assumed constant. Treating the MMSPE estimates of ϕ_δ and τ as its true value, we compute the posterior of $\delta(x)$

using equations (5.12) and (5.13) and display the posterior mean of $\delta(x)$ along with the corresponding 95% prediction interval in Figure 5.3, which shows that the model bias has a large absolute value but small variance when $M_c \in [0.5, 1.1]$, while a small absolute value but large variance when $M_c \notin [0.5, 1.1]$. The reasons for such results can be seen from Figure 5.2, which shows that

- The discrepancies between computer outputs and physical observation are large when $M_c \in [0.5, 1.1]$ and small when $M_c \notin [0.5, 1.1]$, which explains the absolute value of the bias.

- There are more physical observations for $M_c \in [0.5, 1.1]$ than for $M_c \notin [0.5, 1.1]$, which explains the variance of the bias.

Having the posteriors of $Y^m(x)$ and $\delta(x)$, we calculate the prediction of $Y^r(x)$ using equation (5.31). Figure 5.4 displays predictions of $Y^r(\cdot)$ and 95% prediction intervals along with physical observations as circles and computer outputs as upside triangles. Similarly, the variance is smaller where more observations are available and larger where less observations are available. Oberkampf and Barone (2004) fitted a nonlinear model to physical observations. Their results show that the variance is small for small M_c and large for large M_c and approximately equals to zero for $M_c \in [0, 0.3]$. This is due to the nonlinear form they chose.

We use the results for $\delta(x)$ to illustrate the validation of the computer model. The validation based on $Y^r(x)$ can be done similarly. For pure illustration purpose, we set $\Delta_0 = 0.12$. For example, the 95% prediction interval for $\delta(0.65)$ is (-0.0083, -0.1041). According to equations (5.45) and (5.46), we have $\Delta_{min}(0.65) = 0.0083$ and $\Delta_{max}(0.65) = 0.1041$. Since $\Delta_{max}(0.65) < \Delta_0$, we conclude that the computer output at $x = 0.65$ is acceptable. Table 5.9 gives the validation decisions for $x = 0.65$ and another two points, $x = 0.8$ and $x = 1.2$. Same decisions can also made by simply looking at Figure 5.3. Graphically speaking, with a 95% confidence, we

accept the computer model at x if the 95% prediction interval for $\delta(x)$ is located completely within the two boundaries $\delta(x) = -0.12$ and $\delta(x) = 0.12$; we reject the computer model at x if the 95% prediction interval for $\delta(x)$ is complete located completely outside the two boundaries; and we need more physical observations to make a validation decision at x if the 95% prediction interval for $\delta(x)$ is partially located within the two boundaries. With the traditional approach, we reject the computer model at x if the 95% prediction interval for $\delta(x)$ does not contain zero. Therefore, we reject the computer model at $x = 0.65$ and $x = 0.8$ and fail to reject the computer model at $x = 1.2$, which are different from the validation decisions based on Δ_0.

To see if integrating computer outputs and physical observations together improves the prediction of $Y^r(x)$, we repeat the above calculations with $Y^m(x) \equiv 0$. In other words, we use only physical observations now. The MMSPE estimates of ϕ_δ and τ are found to be $\hat{\phi}_\delta = 1.1$ and $\hat{\tau} = 0.01$ respectively. Comparing Figures 5.4 and 5.5 shows that, for this example, including computer outputs improve little on the prediction of $Y^r(x)$. Such a result is due to the following facts: a) this example has 32 physical observations while only 11 computer outputs, b) D_e and D_m span the same range, and c) the discrepancy between computation and experiment is obvious. Therefore, the 11 computer outputs contain little additional information, and using physical observations alone does a fairly good job in prediction. However, often we have only few physical observations and relatively large number of computer outputs. For a better comparison, we randomly partition the physical observations into two parts, a training sample of size 20 and a testing sample of size 12. We estimate the parameters as above using the training sample and calculate the predictions for the testing sample and the corresponding RMSPEs. Table 5.10 contains the average RMSPEs from ten different partitions, showing that using only physical observations gives a RMSPE (0.0786) just slight larger than the RMSPE (0.0756) based on both

computer outputs and physical observations. Table 5.10 also contains the RMSPE (0.0852) calculated using the approach by Kennedy and O'Hagan (2001). It is larger than the RMSPEs from the proposed Bayesian approach using either only physical observations or both physical observations and computer outputs. The predictions based on only computer outputs have the largest RMSPE (0.1344), almost twice of the others.

Table 5.7: The Compressible Shear Layer Example: physical observations

Source	M_c	y^e	Source	M_c	y^e
69	0.992	0.4640	71	0.945	0.4890
65,70	0.059	1.0000	72	0.510	0.9710
65,70	0.342	0.9780	72	0.640	0.7620
65,70	0.428	1.0000	73	0.860	0.5750
65,70	0.476	0.9810	74,75	0.206	0.9850
65,70	0.636	0.7520	74,75	0.455	0.8170
65,70	0.821	0.6010	74,75	0.691	0.5650
65,70	0.928	0.4600	74,75	0.720	0.6330
65,70	1.119	0.4530	76	0.795	0.5020
65,70	1.309	0.4220	74,75	0.862	0.4570
65,70	1.440	0.4400	74,75	0.985	0.4000
71	0.270	1.3500	77	0.525	1.0580
71	0.519	0.9570	77	0.535	0.8100
71	0.589	0.8120	78	0.580	0.9270
71	0.668	0.7330	77	0.640	0.8410
71	0.825	0.5350	78,79	1.040	0.5180

5.6.4 Sensitivity Study

In examples 1 and 2, we have estimated parameters $\boldsymbol{\phi}_\delta$ and τ by maximizing a likelihood function. For example, the ML estimates of $\boldsymbol{\phi}_\delta$ and τ in the fluidized-bed coating example are $\boldsymbol{\phi}_\delta = (0.1518, 0.2833, 1.0274, 0.3726, 0.0, 0.0)^T$ and $\tau = 0.1677$ respectively. We then plug in those estimates into equations (5.12) and (5.13) to get the posterior distribution of the model bias $\delta(\mathbf{x})$. ML estimation involves solving an optimization problem that has an objective function containing the determinant

Table 5.8: The Compressible Shear Layer Example: computer outputs

M_c	d_i	d_c	$y^m = d_c/d_i$
0.100	0.0514	0.0514	1.0000
0.240	0.0711	0.0712	1.0014
0.380	0.0792	0.0760	0.9596
0.520	0.0836	0.0738	0.8828
0.660	0.0865	0.0690	0.7977
0.800	0.0837	0.0630	0.7527
0.940	0.0895	0.0568	0.6346
1.080	0.0898	0.0508	0.5657
1.220	0.0890	0.0455	0.5112
1.360	0.0863	0.0407	0.4716
1.500	0.0810	0.0367	0.4531

Table 5.9: The Compressible Shear Layer Example: model validation with $\Delta_0 = 0.12$

	$x = 0.65$	$x = 0.8$	$x = 1.2$
95% PI for $\delta(x)$	(-0.0083, -0.1041)	(-0.2582, -0.1491)	(-0.3145, 0.0047)
$\Delta_{min}(x)$	0.0083	0.1491	0
$\Delta_{max}(x)$	0.1041	0.2582	0.3145
Condition	$\Delta_{max}(0.65) < \Delta_0$	$\Delta_{min}(0.8) > \Delta_0$	$\Delta_{min}(1.2) < \Delta_0 < \Delta_{max}(1.2)$
Decision	Accept	Reject	Uncertain

Table 5.10: The Compressible Shear Layer Example: RMSPEs

	RMSPE			
Partition	$\hat{y}^r_m = \hat{y}_m$	$\hat{y}^r_e$	$\hat{y}^r$	$\hat{y}^r_{KO}$
1	0.1363	0.0685	0.0669	0.0858
2	0.1314	0.0941	0.0892	0.0990
3	0.1423	0.0872	0.0801	0.1159
4	0.1260	0.0886	0.0852	0.0818
5	0.1478	0.0804	0.0866	0.0938
6	0.1451	0.0849	0.0774	0.0843
7	0.1200	0.0584	0.0542	0.0493
8	0.1179	0.0784	0.0797	0.0815
9	0.1477	0.0738	0.0773	0.0965
10	0.1297	0.0714	0.0624	0.0638
Average	0.1344	0.0786	0.0759	0.0852

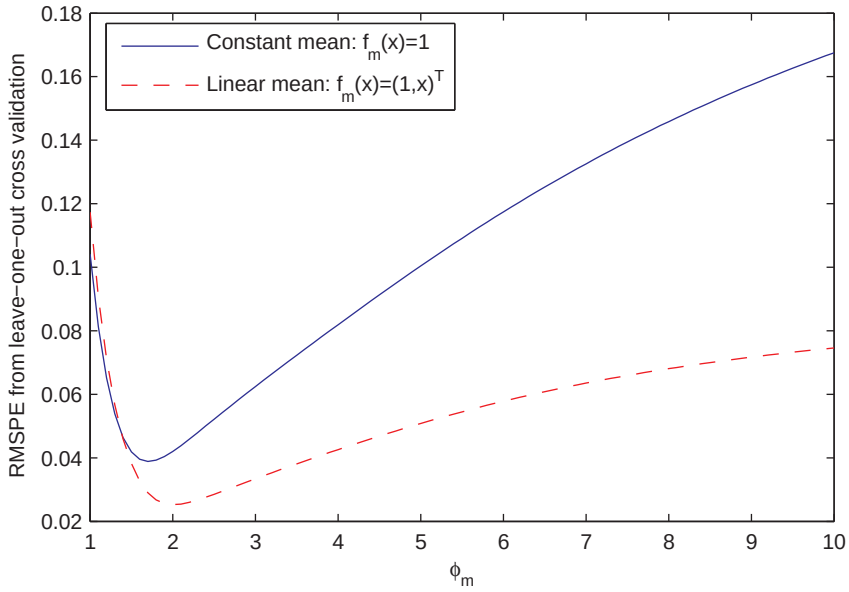


Figure 5.1: The Compressible Shear Layer Example: RMSPE as a function of ϕ_m.

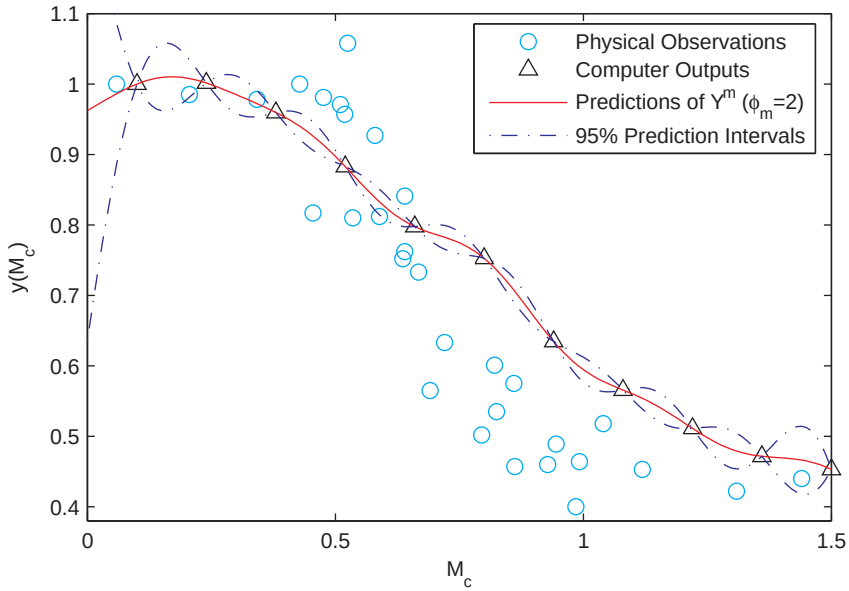


Figure 5.2: The Compressible Shear Layer Example: prediction of $Y^m(\mathbf{x})$.

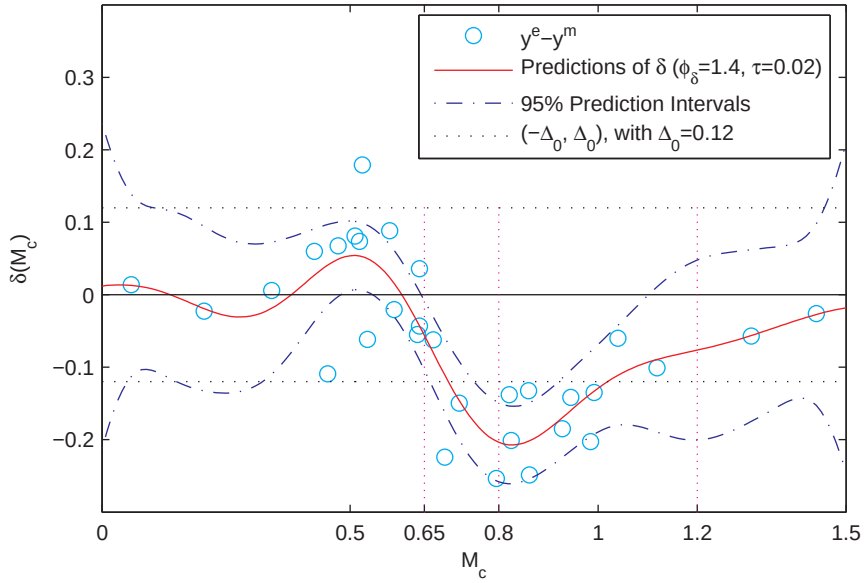


Figure 5.3: The Compressible Shear Layer Example: prediction of $\delta(\mathbf{x})$.

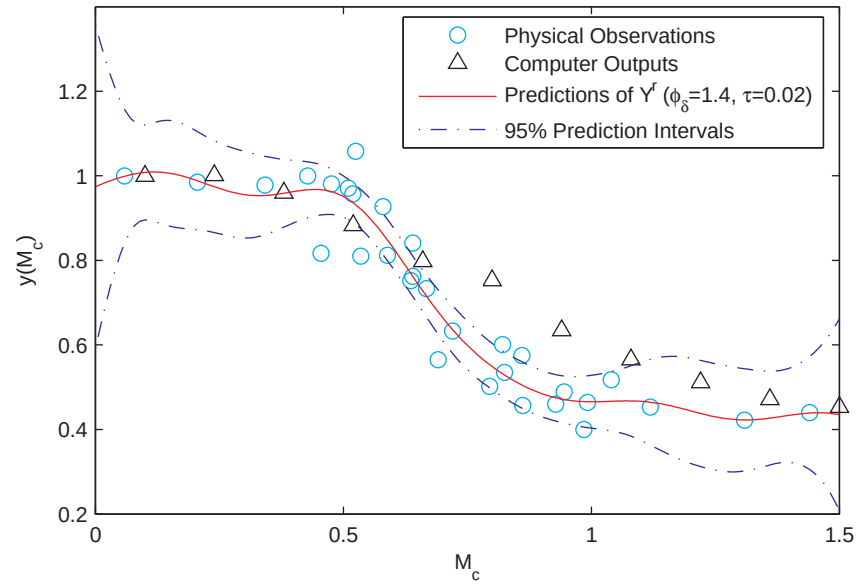


Figure 5.4: The Compressible Shear Layer Example: prediction of $Y^r(\mathbf{x})$ using both computer outputs $\mathbf{y}^m$ and physical observations $\mathbf{y}^e$.

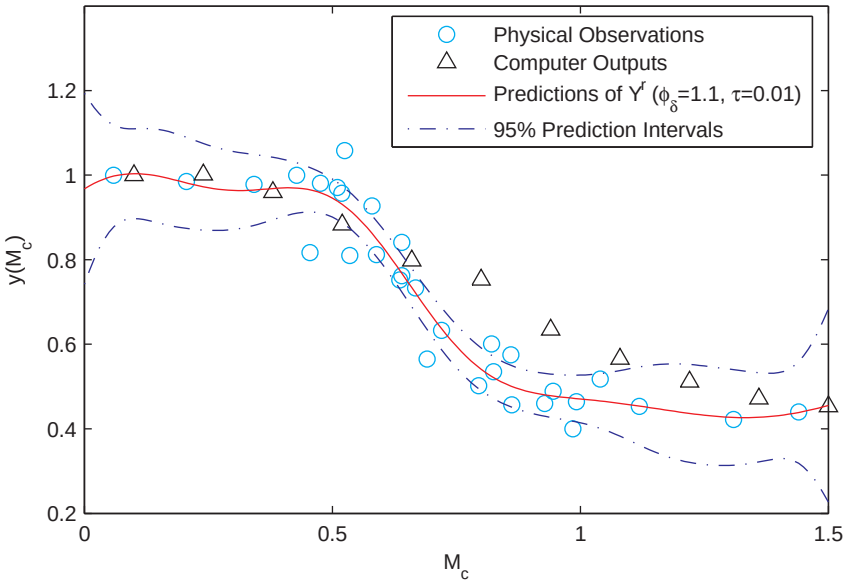


Figure 5.5: The Compressible Shear Layer Example: prediction of $Y^r(\mathbf{x})$ using only physical observations $\mathbf{y}^e$.

of a covariance matrix. This leads to certain numerical difficulties, such as non-convergence and locality, in solving the optimization problem. As a result, we either can not get an optimal solution or have a local optimum. In this section, we study how sensitive the predictions of $\delta(\mathbf{x})$ and therefore $Y^r(\mathbf{x})$ are to the values of $\boldsymbol{\phi}_\delta$ and τ. The purpose is to see how important good estimates of parameters are to the proposed Bayesian approach. If, for example, we find out that the predictions of $\delta(\mathbf{x})$ and $Y^r(\mathbf{x})$ are quite robust to the values of $\boldsymbol{\phi}_\delta$ and τ, then a local optimum might be good enough for the purpose of prediction.

To study the sensitivity of the prediction of $\delta(\mathbf{x})$ to the values of $\boldsymbol{\phi}_\delta$ and τ, we use the fluidized-bed coating example and apply the proposed Bayesian approach with different values of $\boldsymbol{\phi}_\delta$ and τ. For simplicity, we assume that $\phi_{\delta,i} = \phi_\delta$, $1 \le i \le p$, that is, $\boldsymbol{\phi}_\delta = \phi_\delta \cdot (1, \cdots, 1)^T$. Same as in example 1, we reserve runs 4, 15, 17, 21, 23, 25, 26, and 28 as the testing data. We run the proposed Bayesian approach with different values of ϕ_δ and τ and compute the corresponding RMSPEs. The results are displayed in Figures 5.6. Each boxplot in the upper panel of Figure 5.6 corresponds

to a specific value of ϕ_δ and is the boxplot of RMSPEs computed at different values of τ. Similarly, each boxplot in the lower panel of Figure 5.6 corresponds to a specific value of τ and is the boxplot of RMSPEs computed at different values of ϕ_δ.

The upper panel of Figure 5.6 shows that, given the value of ϕ_δ, the value of τ generally has a decreasing effect on RMSPE as the value of ϕ_δ increases (since the length of the boxplot of RMSPE decreases with the increase of the value of ϕ_δ) and has little effect when the value of ϕ_δ is great than 3 (since boxplots for $\phi_\delta > 3$ have a length close to zero). In other words, given the value of ϕ_δ, RMSPE becomes less sensitive to the value of τ as the value of ϕ_δ increases and quite robust to the value of τ when $\phi_\delta > 3$. The lower panel of Figure 5.6 shows that the boxplots of RMSPE for different values of ϕ_δ have around the same lengths and locations and are only sightly shorter and higher for larger values of τ, which suggests that the value of τ has little effect on the sensitivity of RMSPE to the value of ϕ_δ. Also, the long boxplots in the lower panel implies that, given the value of τ, RMSPE is sensitive to the value of ϕ_δ. In addition, comparing the upper and lower panels shows that boxplots in the lower panel have a greater length than those in the upper panel, which suggests that RMSPE is more sensitive to the value of ϕ_δ than to the value of τ.

5.7 Appendix

5.7.1 Posterior Distributions of $\boldsymbol{\beta}_m$ and σ_m^2

The posteriors of $\boldsymbol{\beta}_m$ and σ_m^2 are

$$\boldsymbol{\beta}_m|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\boldsymbol{\beta}_m)} = \boldsymbol{\beta}_m|\mathbf{y}^m, \sigma_m^2, \boldsymbol{\phi}_m \ \sim \ N(\mathbf{A}_m\mathbf{v}_m,\ \sigma_m^2\mathbf{A}_m) \tag{5.49}$$

and

$$\begin{aligned}\sigma_m^2|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\boldsymbol{\beta}_m,\sigma_m^2)} &= \sigma_m^2|\mathbf{y}^m, \boldsymbol{\phi}_m \\ &\sim \ IG(\alpha_m + \frac{n_m}{2},\ \gamma_m + \frac{1}{2}\left[(\mathbf{y}^m)^T\mathbf{R}_m^{-1}\mathbf{y}^m + \mathbf{b}_m^T\mathbf{V}_m^{-1}\mathbf{b}_m - \mathbf{v}_m^T\mathbf{A}_m\mathbf{v}_m\right])\end{aligned} \tag{5.50}$$

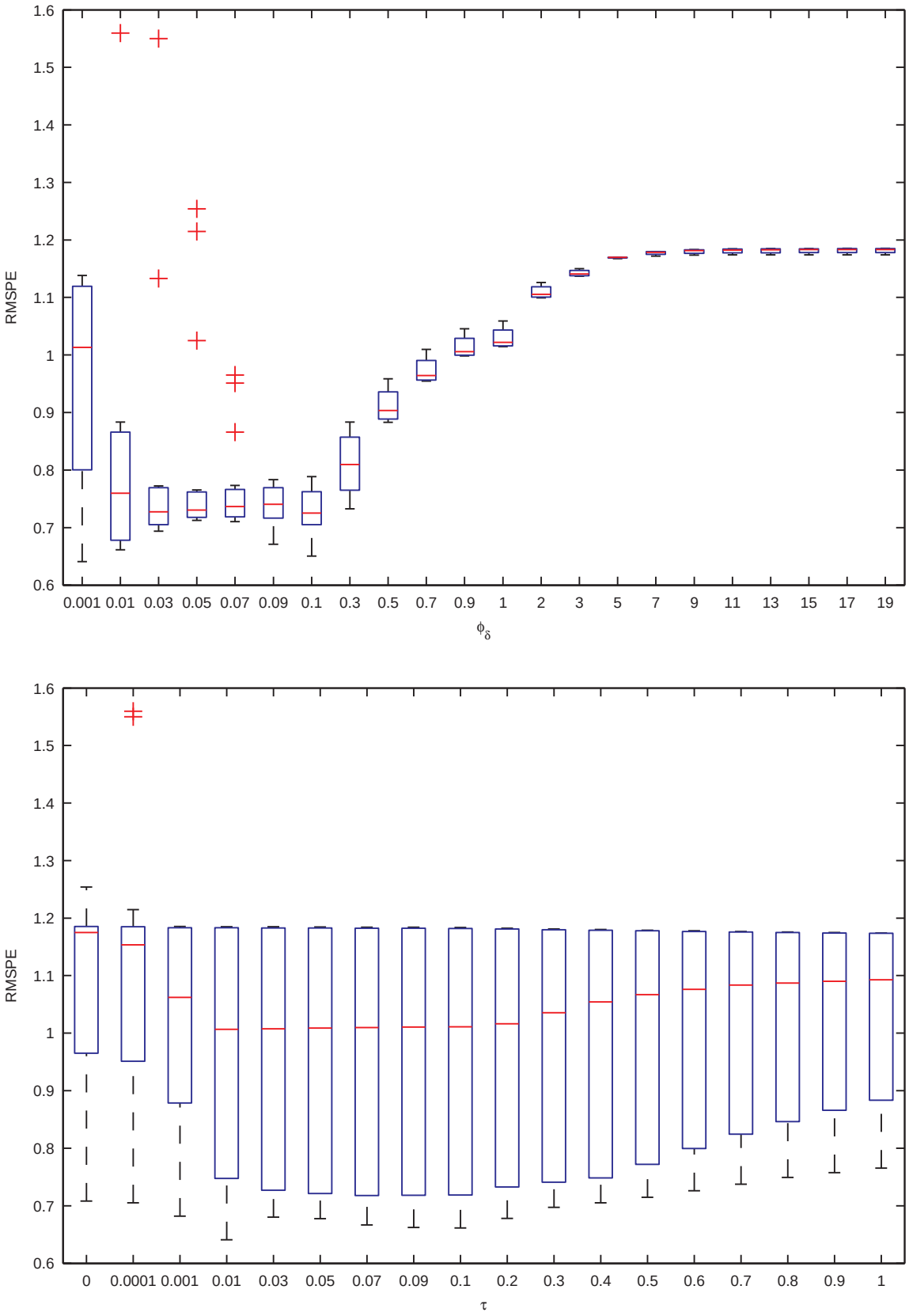


Figure 5.6: The Fluidized-Bed Coating Example: boxplots of RMSPEs

where

$$\mathbf{A}_m^{-1} = \mathbf{F}_m^T \mathbf{R}_m^{-1} \mathbf{F}_m + \mathbf{V}_m^{-1}, \tag{5.51a}$$

$$\mathbf{v}_m = \mathbf{F}_m^T \mathbf{R}_m^{-1} \mathbf{y}^m + \mathbf{V}_m^{-1} \mathbf{b}_m, \tag{5.51b}$$

and $\boldsymbol{\theta}_{-(\cdot)}$ contains all the parameters except those in the parentheses.

5.7.2 Posterior Distributions of $\boldsymbol{\beta}_\delta$ and σ_δ^2

Assume that computer outputs are available at the physical design set D_e. The posteriors of $\boldsymbol{\beta}_\delta$ and σ_δ^2 are

$$\boldsymbol{\beta}_\delta | \mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\boldsymbol{\beta}_\delta)} = \boldsymbol{\beta}_\delta | \mathbf{y}^e, \mathbf{y}_{n_e}^m, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \sigma_\epsilon^2 \ \sim \ N(\mathbf{A}_\delta \mathbf{v}_\delta, \ \sigma_\delta^2 \mathbf{A}_\delta) \tag{5.52}$$

and

$$\begin{aligned} &\sigma_\delta^2 | \mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\boldsymbol{\beta}_\delta, \sigma_\delta^2)} = \sigma_\delta^2 | \mathbf{y}^e, \mathbf{y}_{n_e}^m, \boldsymbol{\phi}_\delta, \tau \\ &\sim \ IG(\alpha_\delta + \frac{n_e}{2}, \ \gamma_\delta + \frac{1}{2} \left[(\mathbf{y}^e - \mathbf{y}_{n_e}^m)^T (\mathbf{R}_\delta(D_e) + \tau \mathbf{I}_{n_e})^{-1} (\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{b}_\delta \mathbf{V}_\delta^{-1} \mathbf{b}_\delta - \mathbf{v}_\delta^T \mathbf{A}_\delta \mathbf{v}_\delta \right]), \end{aligned} \tag{5.53}$$

where

$$\mathbf{A}_\delta^{-1} = \mathbf{F}_\delta^T(D_e)(\mathbf{R}_\delta(D_e) + \tau \mathbf{I}_{n_e})^{-1} \mathbf{F}_\delta(D_e) + \mathbf{V}_\delta^{-1}, \tag{5.54a}$$

$$\mathbf{v}_\delta = \mathbf{F}_\delta^T(D_e)(\mathbf{R}_\delta(D_e) + \tau \mathbf{I}_{n_e})^{-1} (\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{V}_\delta^{-1} \mathbf{b}_\delta, \tag{5.54b}$$

$$\tau = \sigma_\epsilon^2 / \sigma_\delta^2, \tag{5.54c}$$

and $\mathbf{I}_{n_e}$ is an $n_e \times n_e$ identity matrix.

5.7.3 Posterior Distribution of $\delta(D)$

Assuming that $\boldsymbol{\phi}_\delta$ and τ are known, we can rewrite the posterior of $\delta(D)$ in equation (5.21) as

$$\begin{aligned} &p(\delta(D) | \mathbf{y}^e, \mathbf{y}^m) \\ &= \iint_{\boldsymbol{\beta}_\delta, \sigma_\delta^2} p(\delta(D) | \mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \tau) \cdot p(\boldsymbol{\beta}_\delta | \mathbf{y}^e, \mathbf{y}_{n_e}^m, \sigma_\delta^2, R_\delta, \tau) \cdot p(\sigma_\delta^2 | \mathbf{y}^e, \mathbf{y}_{n_e}^m, R_\delta, \tau) \, \mathrm{d}\boldsymbol{\beta}_\delta \, \mathrm{d}\sigma_\delta^2, \end{aligned} \tag{5.55}$$

in which

$$p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \tau) \tag{5.56}$$
$$\propto (\sigma_\delta^2)^{-\frac{n}{2}} \cdot \exp\{-(\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) - \mathbf{H}_\delta^T\boldsymbol{\beta}_\delta)^T$$
$$\cdot [\mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{R}_\delta(D_e, D)]^{-1} \cdot (\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) - \mathbf{H}_\delta^T\boldsymbol{\beta}_\delta)\},$$

$$p(\boldsymbol{\beta}_\delta|\mathbf{y}^e, \mathbf{y}^m, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \tau) \propto (\sigma_\delta^2)^{-\frac{p+1}{2}} \cdot \exp\left\{-\frac{(\boldsymbol{\beta}_\delta - \mathbf{A}_\delta\mathbf{v}_\delta)^T\mathbf{A}_\delta^{-1}(\boldsymbol{\beta}_\delta - \mathbf{A}_\delta\mathbf{v}_\delta)}{2\sigma_\delta^2}\right\}, \tag{5.57}$$

$$p(\sigma_\delta^2|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\phi}_\delta, \tau) \tag{5.58}$$
$$\propto (\sigma_\delta^2)^{-\alpha_\delta - \frac{n_e}{2} - 1} \cdot \exp\left\{-\frac{\gamma_\delta + \frac{1}{2}\left[(\mathbf{y}^e - \mathbf{y}_{n_e}^m)^T\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{b}_\delta\mathbf{V}_\delta^{-1}\mathbf{b}_\delta - \mathbf{v}_\delta^T\mathbf{A}_\delta\mathbf{v}_\delta\right]}{\sigma_\delta^2}\right\},$$

where $\mathbf{B} = (\mathbf{R}_\delta(D_e) + \tau\mathbf{I}_{n_e})^{-1}$, and $\mathbf{H}_\delta^T = \mathbf{F}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{F}_\delta(D_e)$. Collecting all terms involving $\boldsymbol{\beta}_\delta$ together gives

$$\exp\left\{-\frac{1}{2\sigma_\delta^2}\left[\boldsymbol{\beta}_\delta^T\mathbf{A}^{-1}\boldsymbol{\beta}_\delta - 2\mathbf{v}^T\boldsymbol{\beta}_\delta\right]\right\},$$

where

$$\mathbf{A}^{-1} = \mathbf{A}_\delta^{-1} + \mathbf{H}_\delta\left[\mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{R}_\delta(D_e, D)\right]^{-1}\mathbf{H}_\delta^T,$$
$$\mathbf{v} = \mathbf{v}_\delta + \mathbf{H}_\delta\left[\mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{R}_\delta(D_e, D)\right]^{-1}(\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m)).$$

Hence,

$$\int_{\boldsymbol{\beta}_\delta} \exp\left\{-\frac{1}{2\sigma_\delta^2}\left[\boldsymbol{\beta}_\delta^T\mathbf{A}^{-1}\boldsymbol{\beta}_\delta - 2\mathbf{v}^T\boldsymbol{\beta}_\delta\right]\right\} \mathrm{d}\boldsymbol{\beta}_\delta \propto (\sigma_\delta^2)^{\frac{p+1}{2}} \cdot \exp\left\{\frac{\mathbf{v}^T\mathbf{A}\mathbf{v}}{2\sigma_\delta^2}\right\}.$$

Collecting all terms involving σ_δ^2 gives

$$(\sigma_\delta^2)^{-\frac{n}{2} - \alpha_\delta - \frac{n_e}{2} - 1} \cdot \exp\left\{-\frac{\gamma}{\sigma_\delta^2}\right\},$$

where

$$\gamma = \gamma_\delta + \frac{(\mathbf{y}^e - \mathbf{y}_{n_e}^m)^T\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{b}_\delta\mathbf{V}_\delta^{-1}\mathbf{b}_\delta}{2} - \frac{\mathbf{v}^T\mathbf{A}\mathbf{v}}{2}$$
$$+\frac{1}{2}(\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m))^T \cdot [\mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{R}_\delta(D_e, D)]^{-1}$$
$$\cdot(\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m)).$$

Performing the integration over σ_δ^2 yields

$$\int_{\sigma_\delta^2} (\sigma_\delta^2)^{-\frac{n}{2}-\alpha_\delta-\frac{n_e}{2}-1} \exp(-\frac{\gamma}{\sigma_\delta^2})\, \mathrm{d}\sigma_\delta^2 \;\propto\; \gamma^{-\frac{n+2\alpha_\delta+n_e}{2}}.$$

Therefore,

$$p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m) \;\propto\; \gamma^{-\frac{n+2\alpha_\delta+n_e}{2}}.$$

If we can write γ in the form of

$$\gamma = C \cdot \left[1 + \frac{(\delta(D) - \mu_{\delta|e,m}(D))^T \cdot \Sigma_{\delta|e,m}^{-1}(D) \cdot (\delta(D) - \mu_{\delta|e,m}(D))}{2\alpha_\delta + n_e}\right],$$

where C is any constant, then

$$p(\delta(D)|\mathbf{y}^e, \mathbf{y}^m) \;\propto\; \left[1 + \frac{(\delta(D) - \mu_{\delta|e,m}(D))^T \cdot \Sigma_{\delta|e,m}^{-1}(D) \cdot (\delta(D) - \mu_{\delta|e,m}(D))}{2\alpha_\delta + n_e}\right]^{-\frac{n+2\alpha_\delta+n_e}{2}},$$

which implies that $\delta(D)|\mathbf{y}^e, \mathbf{y}^m$ has a multivariate noncentral t distribution with degree of freedom $2\alpha_\delta + n_e$, noncentrality parameter $\mu_{\delta|e,m}(D)$, and scale matrix $\Sigma_{\delta|e,m}(D)$.

Since

$$\mathbf{A} = \mathbf{A}_\delta - \mathbf{A}_\delta\mathbf{H}_\delta \left[\mathbf{H}_\delta^T\mathbf{A}_\delta\mathbf{H}_\delta + \mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{R}_\delta(D_e, D)\right]^{-1} \mathbf{H}_\delta^T\mathbf{A}_\delta,$$

$$\mathbf{v} = \mathbf{v}_\delta + \mathbf{H}_\delta \left[\mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{R}_\delta(D_e, D)\right]^{-1} (\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m)).$$

we have

$$\begin{aligned}\gamma = \gamma_\delta &+ \frac{1}{2}\left[(\mathbf{y}^e - \mathbf{y}_{n_e}^m)^T\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{b}_\delta\mathbf{V}_\delta^{-1}\mathbf{b}_\delta - \mathbf{v}_\delta^T\mathbf{A}_\delta\mathbf{v}_\delta\right] \\ &+\frac{1}{2}(\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) - \mathbf{H}_\delta^T\mathbf{A}_\delta\mathbf{v}_\delta)^T \\ &\quad \cdot \left[\mathbf{H}_\delta^T\mathbf{A}_\delta\mathbf{H}_\delta + \mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{R}_\delta(D_e, D)\right]^{-1} \\ &\quad \cdot(\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) - \mathbf{H}_\delta^T\mathbf{A}_\delta\mathbf{v}_\delta).\end{aligned}$$

Therefore,

$$\mu_{\delta|e,m}(D) = \mathbf{H}_\delta^T \mathbf{A}_\delta \mathbf{v}_\delta + \mathbf{R}_\delta(D, D_e)\mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m)$$
$$\Sigma_{\delta|e,m}(D) = \frac{Q_\delta^2}{2\alpha_\delta + n_e} \cdot [\mathbf{H}_\delta^T \mathbf{A}_\delta \mathbf{H}_\delta + \mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{B}\mathbf{R}_\delta(D_e, D)]$$

where

$$Q_\delta^2 = 2\gamma_\delta + (\mathbf{y}^e - \mathbf{y}_{n_e}^m)^T \mathbf{B}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{b}_\delta \mathbf{V}_\delta^{-1} \mathbf{b}_\delta - \mathbf{v}_\delta^T \mathbf{A}_\delta \mathbf{v}_\delta$$

Substituting $\mathbf{B}$ and $\mathbf{h}$ into above equations for $\mu_{\delta|2}$ and $\sigma_{\delta|2}^2$, we have

$$\mu_{\delta|2}(\mathbf{x}) = \mathbf{f}_\delta^T(D)\mathbf{A}_\delta \mathbf{v}_\delta + \mathbf{r}_\delta^T(D)(\mathbf{R}_\delta + \tau \mathbf{I}_{n_e})^{-1}(\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta \mathbf{A}_\delta \mathbf{v}_\delta)$$
$$\sigma_{\delta|2}^2(\mathbf{x}) = \frac{Q^2}{2\alpha_\delta + n_e} \cdot \Big(1 - \begin{bmatrix} \mathbf{f}_\delta(D) \\ \mathbf{r}_\delta(D) \end{bmatrix}^T \begin{bmatrix} -\mathbf{V}_\delta^{-1} & \mathbf{F}_\delta^T \\ \mathbf{F}_\delta & \mathbf{R}_\delta + \tau \mathbf{I}_{n_e} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}_\delta(D) \\ \mathbf{r}_\delta(D) \end{bmatrix}\Big)$$

where

$$Q^2 = 2\gamma_\delta + (\mathbf{y}^e - \mathbf{y}_{n_e}^m)^T (\mathbf{R}_\delta + \tau \mathbf{I}_{n_e})^{-1}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{b}_\delta \mathbf{V}_\delta^{-1} \mathbf{b}_\delta - \mathbf{v}_\delta^T \mathbf{A}_\delta \mathbf{v}_\delta$$

5.7.4 Densities $p(\mathbf{y}^m|\boldsymbol{\phi}_m)$ and $p(\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\phi}_\delta, \tau)$

The density of $\mathbf{y}^m|\boldsymbol{\phi}_m$ is given by

$$p(\mathbf{y}^m|\boldsymbol{\phi}_m) = \iint_{\boldsymbol{\beta}_m, \sigma_m^2} p(\mathbf{y}^m|\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m) \cdot p(\boldsymbol{\beta}_m|\sigma_m^2) \cdot p(\sigma_m^2)\, \mathrm{d}\boldsymbol{\beta}_m\, \mathrm{d}\sigma_m^2. \tag{5.59}$$

Integrating out $\boldsymbol{\beta}_m$ and σ_m^2 gives

$$p(\mathbf{y}^m|\boldsymbol{\phi}_m) \propto |\mathbf{R}_m(D_m)|^{-\frac{1}{2}} \cdot |\mathbf{A}_m|^{\frac{1}{2}} \tag{5.60}$$
$$\cdot \big[2\gamma_m + (\mathbf{y}^m)^T \mathbf{R}_m^{-1}(D_m)\mathbf{y}^m + \mathbf{b}_m^T \mathbf{V}_m^{-1} \mathbf{b}_m - \mathbf{v}_m^T \mathbf{A}_m \mathbf{v}_m\big]^{-\frac{n_m}{2} - \alpha_m}$$

The density of $by^e|\mathbf{y}^m, \boldsymbol{\phi}_\delta, \sigma_\epsilon^2$ is given by

$$p(\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\phi}_\delta, \tau) \;=\; \iint_{\boldsymbol{\beta}_\delta, \sigma_\delta^2} p(\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, R_\delta, \tau) \cdot p(\boldsymbol{\beta}_\delta|\sigma_\delta^2) \cdot p(\sigma_\delta^2)\, \mathrm{d}\boldsymbol{\beta}_\delta\, \mathrm{d}\sigma_\delta^2 \tag{5.61}$$

Integrating out $\boldsymbol{\beta}_\delta$ and σ_δ^2 gives

$$p(\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\phi}_\delta, \tau) \propto |\mathbf{R}_\delta(D_e) + \tau \mathbf{I}_{n_e}|^{-\frac{1}{2}} \cdot |\mathbf{A}_\delta|^{\frac{1}{2}} \tag{5.62}$$
$$\cdot \big[2\gamma_\delta + (\mathbf{y}^e - \mathbf{y}_{n_e}^m)^T (\mathbf{R}_\delta(D_e) + \tau \mathbf{I}_{n_e})^{-1}(\mathbf{y}^e - \mathbf{y}_{n_e}^m) + \mathbf{b}_\delta^T \mathbf{V}_\delta^{-1} \mathbf{b}_\delta - \mathbf{v}_\delta^T \mathbf{A}_\delta \mathbf{v}_\delta\big]^{-\frac{n_e}{2} - \alpha_\delta}$$

CHAPTER VI

BAYESIAN VALIDATION OF COMPUTER MODELS: PERFORMANCE AND GENERALIZATION

6.1 Introduction

In Chapter 5, we proposed a full Bayesian approach to the validation of computer models. The proposed approach is based on an assumed relationship among the computer model output $Y^m(\mathbf{x})$, the physical observation $Y^e(\mathbf{x})$, and the real system output $Y^r(\mathbf{x})$

$$Y^e(\mathbf{x}) = Y^r(\mathbf{x}) + \epsilon(\mathbf{x}) = Y^m(\mathbf{x}) + \delta(\mathbf{x}) + \epsilon(\mathbf{x}), \tag{6.1}$$

in which the computer model output $Y^m(\mathbf{x})$ and the model bias $\delta(\mathbf{x})$ are assumed to be two mutually independent Gaussian processes, and the measurement error $\epsilon(\mathbf{x})$ is independently and normally distributed and independent of the two Gaussian processes, $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$.

In this chapter, we continue the study of the proposed Bayesian approach. We will focus on two areas:

(1). The performance of the proposed approach. By performance, we mean how well the proposed approach predicts the real system output. In Chapter 5, we have illustrated the performance of the proposed Bayesian approach using three real-life examples. The investigation in this chapter is to understand how the performance of the proposed approach is affected by certain factors, such as the variances of Gaussian processes $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ and the number of

replications in physical experiments. Findings from such investigations could provide valuable insights about how the proposed approach performs and when the proposed approach performs well and why.

(2). A possible generalization to the proposed approach. The proposed approach assumes that the two Gaussian processes $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ are mutually independent. This assumption simplifies the derivation of the posterior distributions of $\delta(\mathbf{x})$ and $Y^r(\mathbf{x})$. With this independence assumption, when $D_e \subseteq D_m$ (D_e and D_m are the design sets of the physical and computer experiments respectively), the posterior distributions of $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ given physical observation $\mathbf{y}^e$ at D_e and computer outputs $\mathbf{y}^m$ at D_m are also mutually independent. As a result, the posterior of $Y^r(\mathbf{x})$ can be obtained nicely as the sum of two independent random process. However, the assumption of independence between $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ may not hold in reality. For example, it is possible that the model bias $\delta(\mathbf{x})$ is large where the computer output $Y^m(\mathbf{x})$ is large and small where $Y^m(\mathbf{x})$ is small. In this chapter, we explore the posterior behaviors of $\delta(\mathbf{x})$ and $Y^r(\mathbf{x})$ given $\mathbf{y}^m$ and $\mathbf{y}^e$ when $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ in equation (6.1) are assumed to be correlated.

This chapter is organized as follows. Section 6.2 investigates the performance of the proposed approach under different situations, such as different number of replications in physical experiments. Section 6.3 derives the posterior distributions of $\delta(\mathbf{x})$ and $Y^r(\mathbf{x})$ given $\mathbf{y}^m$ and $\mathbf{y}^e$ when $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ are assumed to be correlated.

6.2 Performance of The Proposed Bayesian Approach

The performance investigation is to understand how the performance of the proposed approach is affected by three factors, the number of replications in physical experiments, the variance of the model bias $\delta(x)$, and the variance of the computer model

$Y^m(\mathbf{x})$. Through this investigation, we hope to have a better understanding on when the proposed approach performs well and why. Such information could be useful in choosing computer and physical designs for computer model validation.

6.2.1 Number of Replications in Physical Experiments

Example 1: We use a simulated example to investigate the influences of the number of replications in physical experiments on the prediction of the real system output $Y^r(x)$. This example has one input variable x taking values in the interval $[0, 10]$. The model bias $\delta(x)$ shown in Figure 6.1 is generated as one realization of a Gaussian process with $\mu_\delta(x) = 0.2x$, $\sigma_\delta^2 = 1$, $\phi_\delta = 1$, and $P_\delta = 2$. The computer model $Y^m(x)$ shown in Figure 6.2 is generated as one realization of a Gaussian process with $\mu_m(x) = 10$, $\sigma_m^2 = 1$, $\phi_m = 2$, and $P_m = 2$. The real system output $Y^r(x)$ is calculated as the sum of $Y^m(x)$ and $\delta(x)$.

The $Y^m(x)$, $\delta(x)$, and $Y^r(x)$ generated above are the true computer model, model bias, and system output, which we will predict based on data collected from computer and physical experiments. We now simulate computer and physical experiments, that is, determine computer design set D_m, computer outputs $\mathbf{y}^m$ at D_m, physical design set D_e, and physical observations $\mathbf{y}^e$ at D_e. The computer design is given in Table 6.1 with design set $D_m = \{x_1^m, \cdots, x_{20}^m\}$. At each $x_i^m \in D_m$, $i = 1, \cdots, 20$, we compute the corresponding computer output $y^m(x_i^m)$ as $Y^m(x_i^m)$, the output of the generated computer model. We then have the vector of computer outputs $\mathbf{y}^m = (y^m(x_1^m), \cdots, y^m(x_{20}^m))^T$. The physical design is given in Table 6.2 with design set $D_e = \{x_1^e, \cdots, x_7^e\}$. At each $x_i^e \in D_e$, $i = 1, \cdots, 7$, we compute the corresponding physical observation $y^e(x_{ij}^e)$ as $Y^r(x_i^e)(= Y^m(x_i^e) + \delta(x_i^e))$ plus an error term $\epsilon(x_{ij}^e)$ generated from a normal distribution with mean zero and variance $\sigma_\epsilon^2 = 1$, where $j = 1, \cdots, J$, and J is the number of replications. Therefore, the vector of physical observations $\mathbf{y}^e = (y^e(x_{11}^e), \cdots, y^e(x_{1J}^e), \cdots, y^e(x_{71}^e), \cdots, y^e(x_{7J}^e))^T$. In our study, the

number of replications J is set to be 1, 2, 5, 10, and 20. Figure 6.3 illustrates computer outputs $\mathbf{y}^m$ and physical observations $\mathbf{y}^e$ together with the computer model $Y^m(x)$ and the real system output $Y^r(x)$ for the number of replications $J = 10$.

We use D_m, $\mathbf{y}^m$, D_e, $\mathbf{y}^e$ as the training data to estimate the posterior distributions of $\delta(x)$, $Y^m(x)$, and $Y^r(x)$, and then predict the real system output using its estimated posterior mean. Table 6.3 gives the RMSPEs of the predictions of $Y^r(x)$ at 201 x values from 0 to 10 with an increment 0.05. In this table, $\hat{Y}^r_m$, $\hat{Y}^r_e$, and $\hat{Y}^r$ denote the predictions of $Y^r(x)$ based on only $\mathbf{y}^m$, only $\mathbf{y}^e$, and both $\mathbf{y}^m$ and $\mathbf{y}^e$ respectively. Clearly, using both $\mathbf{y}^m$ and $\mathbf{y}^e$ leads to smaller RMSPEs than using either only $\mathbf{y}^e$ or $\mathbf{y}^m$. Furthermore, the RMSPEs based on $\mathbf{y}^e$ (considering $\mathbf{y}^m$ or not) decrease as the number of replications J increases. Figures 6.4 – 6.8 display the predictions of $Y^r(x)$ based on both $\mathbf{y}^m$ and $\mathbf{y}^e$ and the corresponding 95% confidence intervals for $J = 1, 2, 5, 10$, and 20 respectively. Also shown in those figures are the real system output $Y^r(x)$ and the predictions of $Y^r(x)$ based on only $\mathbf{y}^e$. From Figures 6.4 – 6.8, it is clear that the predictions based on both $\mathbf{y}^m$ and $\mathbf{y}^e$ are closer to $Y^r(x)$ than those based on only $\mathbf{y}^e$. Furthermore, the 95% confidence intervals for $Y^r(x)$ become narrower as the number of replications increases especially at those points where physical observations are available (i.e., those points in the physical design set D_e). This can be seen more clearly from Figure 6.9, which shows that the estimated variances of $Y^r(x)$ decrease as the number of replications increases and decrease faster at those points in D_e. Table 6.4 gives the estimated variances of $Y^r(x)$ at D_e, showing that the estimated variances of $Y^r(x)$ at D_e approach to σ^2_ϵ/J, the variance of the experimental error $\epsilon(x)$ over the number of replications J. In addition, Figures 6.4 – 6.8 show that the choice of the physical design set D_e affects the accuracy of predictions. The predictions of $Y^r(x)$ are closer to the corresponding true values over the region (say $x > 3$) where design points are chosen to capture the major characteristics (such as maxima and minima) of $Y^r(x)$. Table 6.5 gives the estimates

of σ_ϵ^2 based on only $\mathbf{y}^e$ and both $\mathbf{y}^m$ and $\mathbf{y}^e$. This table shows that the estimates of σ_ϵ^2 approach to its true value 1 as the number of replications increases. Moreover, using both $\mathbf{y}^m$ and $\mathbf{y}^e$ leads to a more accurate estimate of σ_ϵ^2 than using only $\mathbf{y}^e$ although the difference between the two estimates becomes smaller as the number of replications gets larger.

The last column of Table 6.3 gives the RMSPEs of the predictions of $Y^m(x)$ at 201 x points using computer outputs $\mathbf{y}^m$. The small RMSPEs imply that, with computer outputs at the chosen 20 computer design points, we can predict computer outputs fairly well. This can also be seen from Figure 6.10 showing that the predictions of $Y^m(x)$ are fairly close to the corresponding true computer outputs and the 95% confidence intervals for $Y^m(x)$ are rather small over almost the entire input region except in a small neighborhood of $x = 0$. Figure 6.11 displays the estimated variances of $Y^m(x)$ that equal to zero at points in D_m and are very small (less than 0.065) at other points except in a small neighborhood of $x = 0$.

We also run the proposed Bayesian procedure using the mean of physical observations at each design point x_i^e in D_e instead of all replications. Tables 6.6 and 6.7 give the RMSPEs of the predictions of $Y^r(x)$ at 201 x values and the estimates of σ_ϵ^2 respectively. The RMSPEs are slightly larger than the RMSPEs based on all replications in Table 6.3, while the estimates of σ_ϵ^2 are far worse than the estimates based on all replications in Table 6.5. In other words, using the means of physical observations leads to a slightly less accurate prediction of $Y^r(x)$ in terms of RMSPE while a much worse estimate of σ_ϵ^2. This can be explained by the fact that the means do not retain all information contained in physical observations especially the information on the variability of physical observations. Figures 6.12 and 6.13 compare the predictions and estimated variances of $Y^r(x)$ based on the means to those based on all replications for $J = 20$. Figure 6.12 suggests that the predictions based on the means are quite close to those based on all replications while Figure 6.13 shows an interesting

phenomenon that the estimated variances based on the means are greater than those based on all replications at points close to physical design points while smaller at points further away from physical design points.

Table 6.1: Computer Outputs at $D_m = \{x_1^m, \cdots, x_{20}^m\}$

Run i	x_i^m	$y^m(x_i^m)$	Run i	x_i^m	$y^m(x_i^m)$
1	0.45	9.3162	11	5.45	11.8907
2	0.95	10.0691	12	5.95	11.9983
3	1.45	9.8351	13	6.45	11.1016
4	1.95	10.2033	14	6.95	10.7167
5	2.45	11.0212	15	7.45	10.0192
6	2.95	11.4428	16	7.95	8.6365
7	3.45	8.9847	17	8.45	9.2830
8	3.95	8.8907	18	8.95	9.7437
9	4.45	9.1616	19	9.45	9.5979
10	4.95	10.4087	20	9.95	9.5222

Table 6.2: Physical Observations at $D_e = \{x_1^e, \cdots, x_7^e\}$ for $J = 1$

Run i	x_i^e	$y^e(x_i^e)$
1	0.45	12.3043
2	1.95	9.0834
3	3.45	12.3266
4	4.95	12.1057
5	6.45	14.1211
6	7.95	8.1999
7	9.45	12.0574

Table 6.3: RMSPEs of Predictions of $Y^r(x)$ or $Y^m(x)$ at 201 x values (from 0 to 10 with an increment 0.05)

	RMSPE			
J	$\hat{Y}^r_m = E[Y^m\|\mathbf{y}^m]$	$\hat{Y}^r_e = E[Y^r\|\mathbf{y}^e]$	$\hat{Y}^r = E[Y^r\|\mathbf{y}^e, \mathbf{y}^m]$	$\hat{Y}^m_m = E[Y^m\|\mathbf{y}^m]$
1	1.6354	1.5111	0.9451	0.1071
2	1.6354	1.5185	0.8231	0.1071
5	1.6354	1.3736	0.7293	0.1071
10	1.6354	1.2790	0.7129	0.1071
20	1.6354	1.2350	0.7017	0.1071

Table 6.4: Estimated $\text{Var}(Y^r(x_i^e)|\mathbf{y}^e, \mathbf{y}^m)$ at $x_i^e \in D_e$, $i = 1, 2, \cdots, 7$

		$\hat{\text{Var}}(Y^r(x_i^e)\|\mathbf{y}^e, \mathbf{y}^m)$						
J	σ^2_ϵ/J	1	2	3	4	5	6	7
1	1.00	1.7892	1.2181	0.8755	0.7612	0.8755	1.2181	1.7892
2	0.50	0.6031	0.4623	0.3778	0.3497	0.3778	0.4623	0.6031
5	0.20	0.2375	0.2109	0.1949	0.1896	0.1949	0.2109	0.2375
10	0.10	0.1088	0.1028	0.0993	0.0981	0.0993	0.1028	0.1088
20	0.05	0.0537	0.0522	0.0513	0.0510	0.0513	0.0522	0.0537

Table 6.5: Estimated Experimental Error Variance σ^2_ϵ

	Using only $\mathbf{y}^e$			Using both $\mathbf{y}^e$ and $\mathbf{y}^m$				
J	$\hat{\sigma}^2_r$	$\hat{\tau}$	$\hat{\sigma}^2_\epsilon$ $(\hat{\tau} \cdot \hat{\sigma}^2_r)$	$\hat{\sigma}^2_m$	$\hat{\sigma}^2_\delta$	$\hat{\sigma}^2_r$ $(\hat{\sigma}^2_m + \hat{\sigma}^2_\delta)$	$\hat{\tau}$	$\hat{\sigma}^2_\epsilon$ $(\hat{\tau} \cdot \hat{\sigma}^2_\delta)$
1	0.4074	14.3467	**5.8449**	0.9322	0.4074	1.3396	10.5076	**4.2809**
2	0.3748	7.5679	**2.8364**	0.9322	0.3468	1.2790	6.6522	**2.3072**
5	0.9059	1.9197	**1.7391**	0.9322	0.3772	1.3094	4.4357	**1.6733**
10	1.0508	1.2459	**1.3092**	0.9322	0.3647	1.2969	3.5680	**1.3014**
20	0.9676	1.2178	**1.1783**	0.9322	0.3606	1.2927	3.2637	**1.1768**

Table 6.6: RMSPEs (using the means of physical observations at D_e)

	RMSPE			
J	$\hat{Y}_m^r = E[Y^m \mid \mathbf{y}^m]$	$\hat{Y}_e^r = E[Y^r \mid \mathbf{y}^e]$	$\hat{Y}^r = E[Y^r \mid \mathbf{y}^e, \mathbf{y}^m]$	$\hat{Y}_m^m = E[Y^m \mid \mathbf{y}^m]$
1	1.6354	1.5111	0.9451	0.1071
2	1.6354	1.5214	0.8158	0.1071
5	1.6354	1.4833	0.7294	0.1071
10	1.6354	1.4632	0.7136	0.1071
20	1.6354	1.4287	0.7047	0.1071

Table 6.7: Estimated Experimental Error Variance σ_ϵ^2 (using the means of physical observations at D_e)

	Using only $\mathbf{y}^e$			Using both $\mathbf{y}^e$ and $\mathbf{y}^m$				
J	$\hat{\sigma}_r^2$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_r^2 \cdot J)$	$\hat{\sigma}_m^2$	$\hat{\sigma}_\delta^2$	$\hat{\sigma}_r^2$ $(\hat{\sigma}_m^2 + \hat{\sigma}_\delta^2)$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_\delta^2 \cdot J)$
1	0.4074	14.3467	**5.8449**	0.9322	0.4074	1.3396	10.5076	**4.2809**
2	0.4074	3.7662	**3.0688**	0.9322	0.4074	1.3396	1.3142	**1.0709**
5	0.4074	6.1959	**12.6213**	0.9322	0.4074	1.3396	1.3412	**2.7320**
10	0.4074	5.5868	**22.7610**	0.9322	0.4074	1.3396	0.5874	**2.3931**
20	0.4074	4.5522	**37.0915**	0.9322	0.4074	1.3396	0.3490	**2.8433**

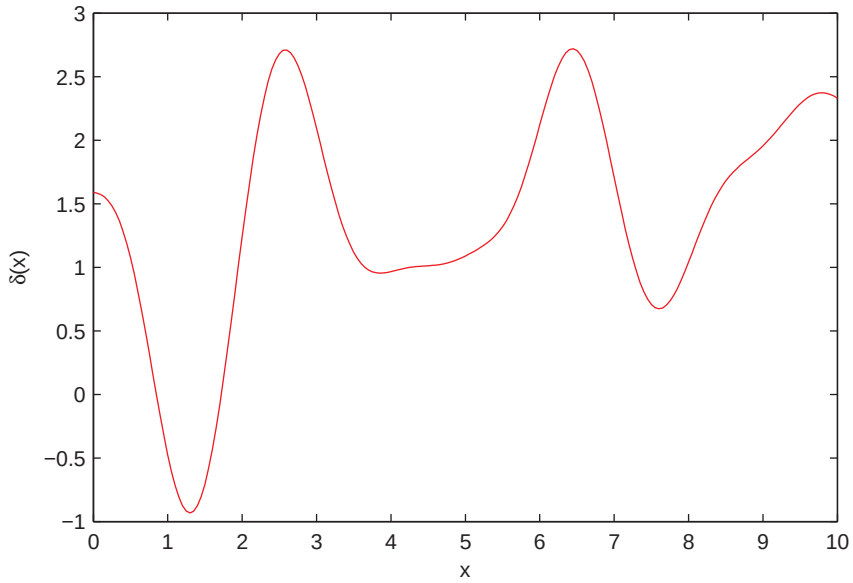


Figure 6.1: Model Bias $\delta(x)$ – one realization of the Gaussian process with $\mu_\delta(x) = 0.2x$, $\sigma_\delta^2 = 1$, $\phi_\delta = 1$, and $P_\delta = 2$

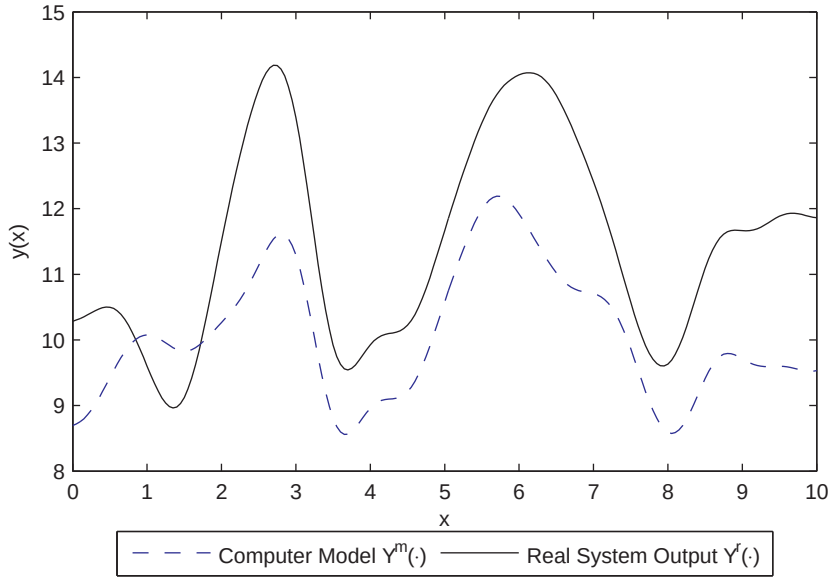


Figure 6.2: Computer Model $Y^m(x)$ – one realization of the Gaussian process with $\mu_m(x) = 10$, $\sigma_m^2 = 1$, $\phi_m = 2$, and $P_m = 2$; Real System Output $Y^r(x) = Y^m(x) + \delta(x)$

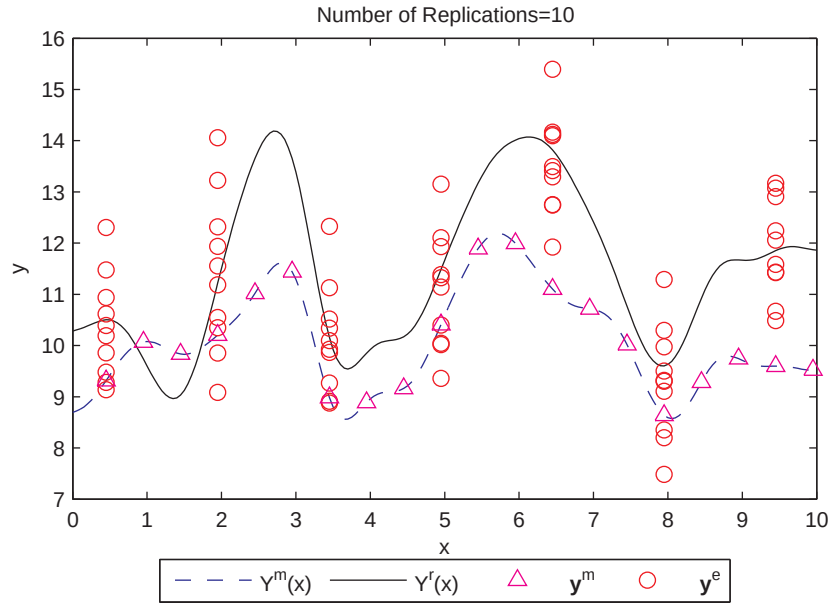


Figure 6.3: Physical Observations $\mathbf{y}^e$ and Computer Outputs $\mathbf{y}^e$ for $J = 10$

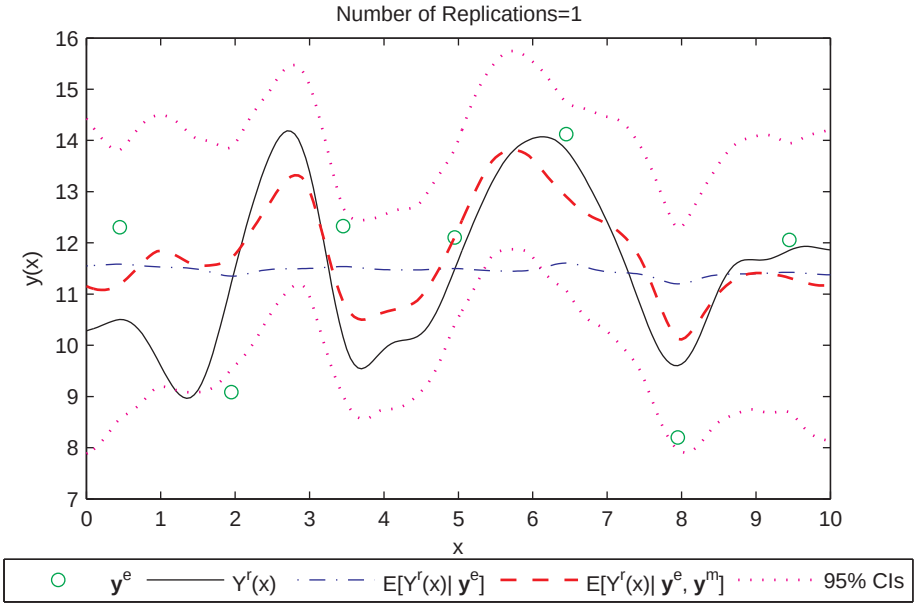


Figure 6.4: Predictions of $Y^r(x)$ for $J = 1$

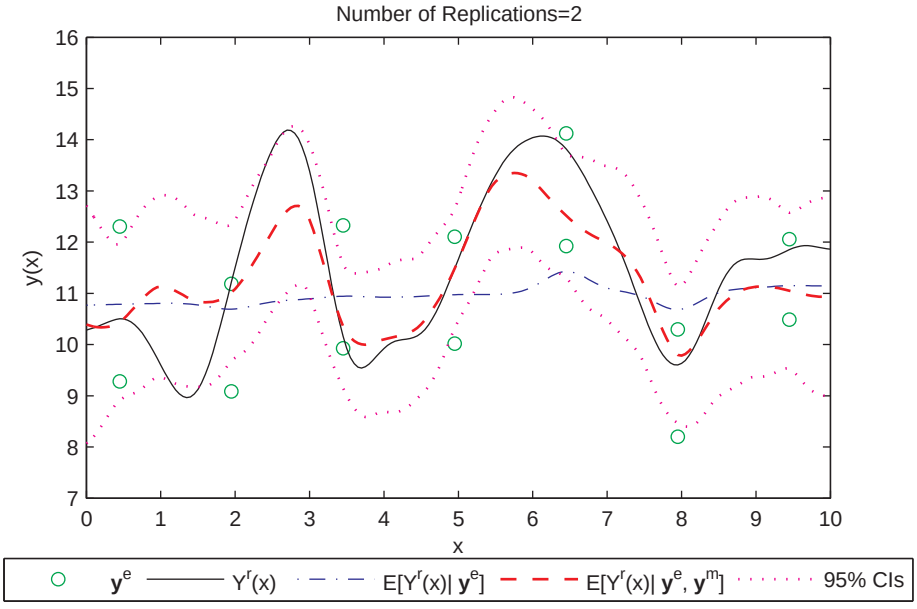


Figure 6.5: Predictions of $Y^r(x)$ for $J = 2$

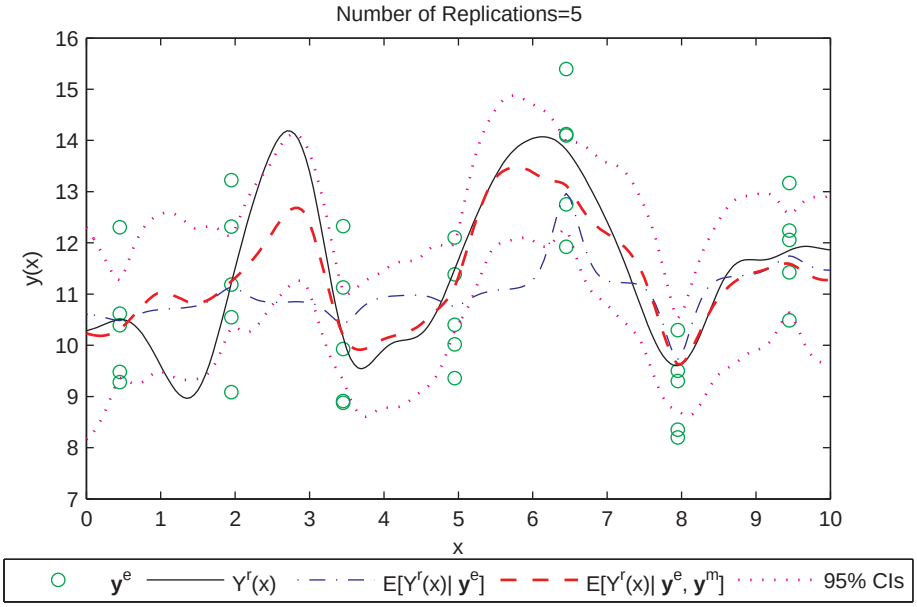


Figure 6.6: Predictions of $Y^r(x)$ for $J = 5$

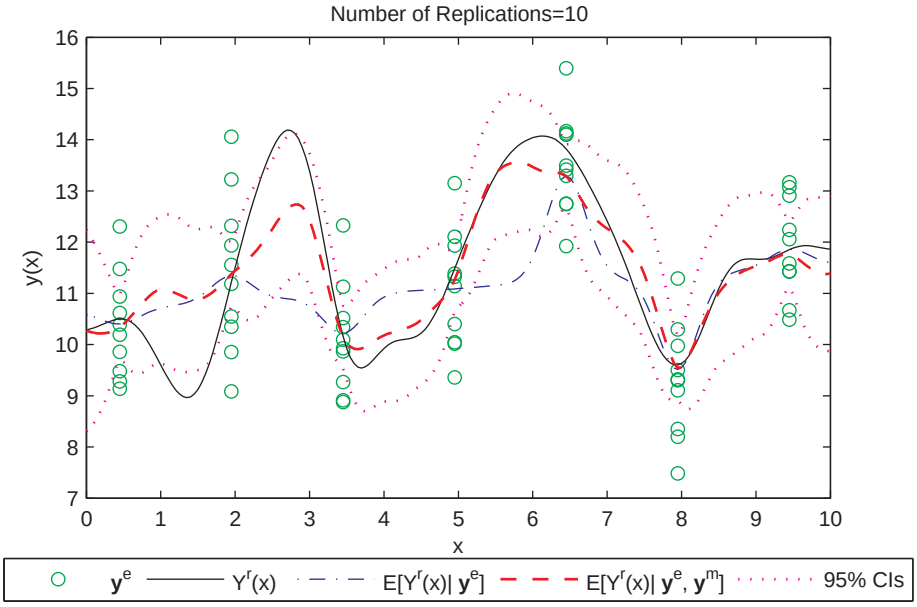


Figure 6.7: Predictions of $Y^r(x)$ for $J = 10$

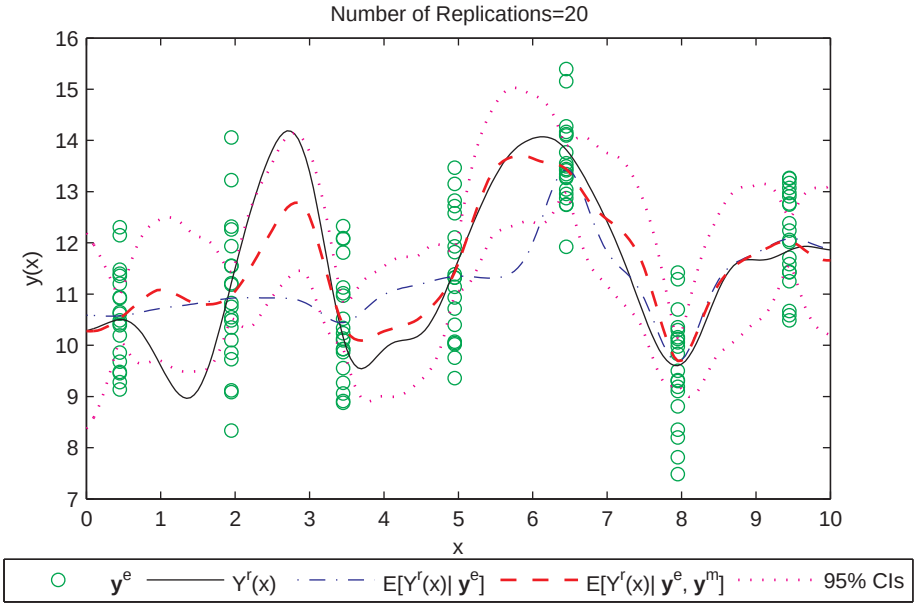


Figure 6.8: Predictions of $Y^r(x)$ for $J = 20$

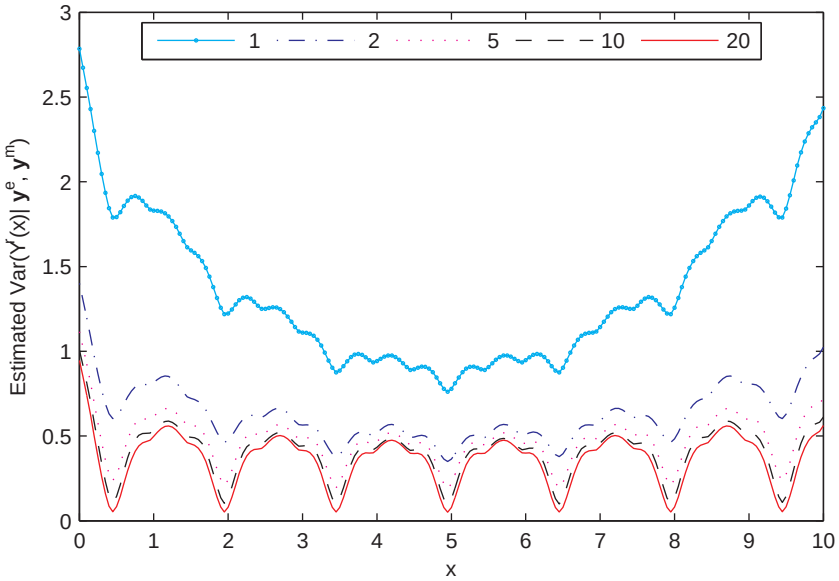


Figure 6.9: Estimated $\mathrm{Var}(Y^r(x)|\mathbf{y}^e, \mathbf{y}^m)$

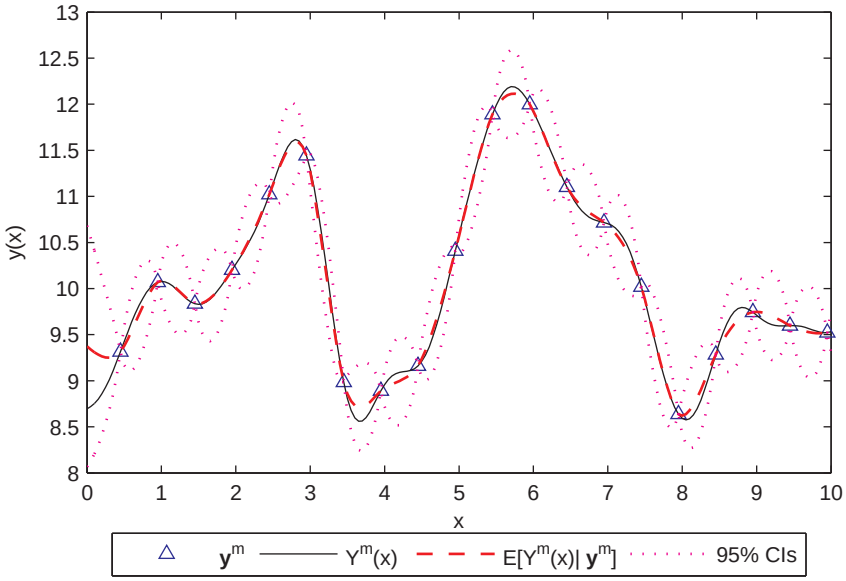


Figure 6.10: Predictions of $Y^m(x)$

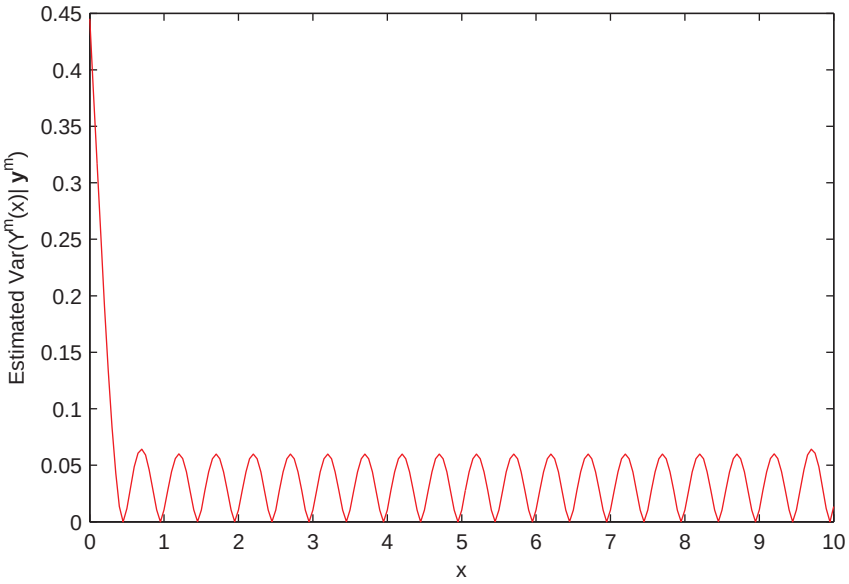


Figure 6.11: Estimated $\mathrm{Var}(Y^m(x)|\mathbf{y}^m)$

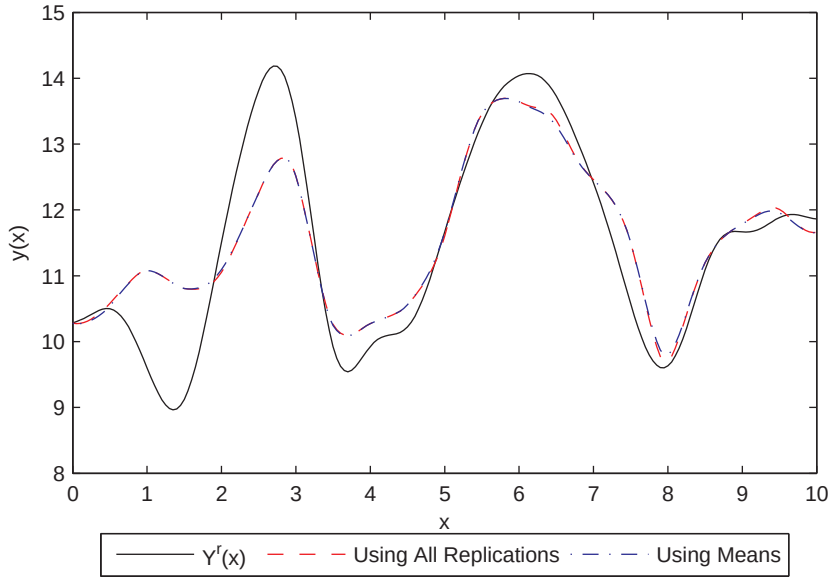


Figure 6.12: Predictions of $Y^r(x)$

6.2.2 Variance of Model Bias $\delta(x)$, σ_δ^2

Example 2: Example 2 is different from Example 1 in only one aspect: the model bias $\delta(x)$ is generated as one realization of a Gaussian process with $\mu_\delta(x) = 0.2x$, $\phi_\delta = 1$, $P_\delta = 2$, and $\sigma_\delta^2 = [0.01, 0.2, 0.5, 1, 2, 5, 10, 20]$. Figure 6.14 displays model biases for different values of σ_δ^2. The purpose here is to study the effects of the value of σ_δ^2 on the prediction of $Y^r(x)$ and the estimation of σ_ϵ^2. We run the proposed Bayesian procedure with the number of replication J as two and five.

Table 6.8 contains the RMSPEs of the predictions of $Y^r(x)$ at 201 x values from 0 to 10 with an increment 0.05. The RMSPEs considering $\mathbf{y}^e$ are smaller for a larger number of replications. Moreover, the RMSPEs based on both $\mathbf{y}^m$ and $\mathbf{y}^e$ are smaller than those based on only $\mathbf{y}^e$ or $\mathbf{y}^m$. Those findings are consistent with the results obtained in Example 1. Table 6.8 also shows that the smaller the value of σ_δ^2, the more accurate the predictions of $Y^r(x)$ in terms of RMSPE. Such a result is expected, and the reason is given as follows. As shown in Figure 6.14, the curvature of $\delta(x)$ increases as the value of σ_δ^2 increases from 0.01 to 20. Therefore, the curvature of the

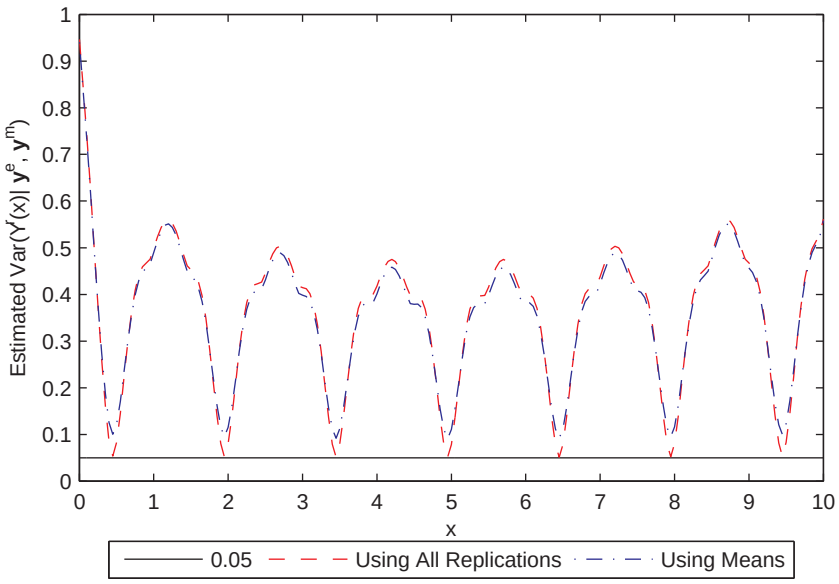


Figure 6.13: Estimated $\mathrm{Var}(Y^r(x)|\mathbf{y}^e, \mathbf{y}^m)$

real system output $Y^r(x)$, dominated by the computer model $Y^m(x)$ when $\sigma_\delta^2 = 0.01$ (see Figure 6.16), becomes more and more dominated by the model bias $\delta(x)$ as the value of σ_δ^2 increases (see Figure 6.17). Since, with the same training data, a curve (or a surface) having higher curvature is more difficult to estimate than one having lower curvature, we predict $\delta(x)$ therefore $Y^r(x)$ better when the value of σ_δ^2 is smaller.

Table 6.9 and Figure 6.15 display the estimates of σ_ϵ^2 for different values of σ_δ^2. Both suggest that the estimates of σ_ϵ^2 are closer to its true value 1 when the number of replications is larger and/or when both $\mathbf{y}^m$ and $\mathbf{y}^e$ are considered, which is consistent with the results obtained in Example 1. Moreover, with the number of replications as two, the estimate of σ_ϵ^2 experiences a very small increase as the value of σ_δ^2 changes from 0.01 to 1 while increases rapidly with the value of σ_δ^2 when it is greater than 1; also, the smaller the value of σ_δ^2, the closer the estimate of σ_ϵ^2 to its true value 1. With the number of replications as five, the estimate of σ_ϵ^2 experiences only a small change as the value of σ_δ^2 increases from 0.01 to 20. In summary, the estimate of σ_ϵ^2 is quite robust to the value of σ_δ^2 when the number of replication is large or the value of σ_δ^2 is small.

Table 6.8: RMSPEs of Predictions of $Y^r(x)$ or $Y^m(x)$ at 201 x values (from 0 to 10 with an increment 0.05)

Number of Replications = 2

		RMSPE			
σ_δ^2	τ	$\hat{Y}_m^r = E[Y^m\|\mathbf{y}^m]$	$\hat{Y}_e^r = E[Y^r\|\mathbf{y}^e]$	$\hat{Y}^r = E[Y^r\|\mathbf{y}^e, \mathbf{y}^m]$	$\hat{Y}_m^m = E[Y^m\|\mathbf{y}^m]$
0.01	100	1.1865	1.1953	0.3721	0.1071
0.2	5	1.3165	1.3131	0.4903	0.1071
0.5	2	1.4496	1.3744	0.6336	0.1071
1	1	1.6260	1.5185	0.8231	0.1071
2	0.5	1.9084	1.7623	1.1153	0.1071
5	0.2	2.5346	2.3098	1.7239	0.1071
10	0.1	3.2919	2.9626	2.4402	0.1071
20	0.05	4.4031	4.0117	3.4854	0.1071

Number of Replications = 5

		RMSPE			
σ_δ^2	τ	$\hat{Y}_m^r = E[Y^m\|\mathbf{y}^m]$	$\hat{Y}_e^r = E[Y^r\|\mathbf{y}^e]$	$\hat{Y}^r = E[Y^r\|\mathbf{y}^e, \mathbf{y}^m]$	$\hat{Y}_m^m = E[Y^m\|\mathbf{y}^m]$
0.01	100	1.1865	1.0265	0.2360	0.1071
0.2	5	1.3165	1.1104	0.3690	0.1071
0.5	2	1.4496	1.2053	0.5276	0.1071
1	1	1.6260	1.3736	0.7293	0.1071
2	0.5	1.9084	1.5882	1.0054	0.1071
5	0.2	2.5346	2.3206	1.7025	0.1071
10	0.1	3.2919	2.9847	2.2159	0.1071
20	0.05	4.4031	3.9613	3.1814	0.1071

Table 6.9: Estimated Experimental Error Variance σ_ϵ^2

Number of Replications = 2

	Using only $\mathbf{y}^e$			Using both $\mathbf{y}^e$ and $\mathbf{y}^m$				
σ_δ^2	$\hat{\sigma}_r^2$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_r^2)$	$\hat{\sigma}_m^2$	$\hat{\sigma}_\delta^2$	$\hat{\sigma}_r^2$ $(\hat{\sigma}_m^2 + \hat{\sigma}_\delta^2)$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_\delta^2)$
0.01	0.3750	5.8942	**2.2103**	0.9322	0.3750	1.3072	5.6856	**2.1321**
0.2	0.3750	6.6480	**2.4930**	0.9322	0.3362	1.2684	6.4156	**2.1569**
0.5	0.3624	7.0904	**2.5694**	0.9322	0.3399	1.2721	6.4930	**2.2068**
1	0.3748	7.5679	**2.8364**	0.9322	0.3468	1.2790	6.6522	**2.3072**
2	0.3906	8.5170	**3.3269**	0.9322	0.3605	1.2927	7.0298	**2.5342**
5	0.4079	11.6682	**4.7591**	0.9322	0.3907	1.3229	8.5220	**3.3294**
10	0.4077	17.5744	**7.1645**	0.9322	0.4084	1.3406	12.0365	**4.9158**
20	0.1532	305.1512	**46.7497**	0.9322	0.4048	1.3370	21.1517	**8.5631**

Number of Replications = 5

	Using only $\mathbf{y}^e$			Using both $\mathbf{y}^e$ and $\mathbf{y}^m$				
σ_δ^2	$\hat{\sigma}_r^2$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_r^2)$	$\hat{\sigma}_m^2$	$\hat{\sigma}_\delta^2$	$\hat{\sigma}_r^2$ $(\hat{\sigma}_m^2 + \hat{\sigma}_\delta^2)$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_\delta^2)$
0.01	0.4346	3.9268	**1.7065**	0.9322	0.3514	1.2835	4.3376	**1.5240**
0.2	0.5582	3.1126	**1.7374**	0.9322	0.2960	1.2282	5.3494	**1.5837**
0.5	0.6970	2.5027	**1.7443**	0.9322	0.3254	1.2575	4.9905	**1.6238**
1	0.9059	1.9197	**1.7391**	0.9322	0.3772	1.3094	4.4357	**1.6733**
2	1.2882	1.3378	**1.7234**	0.9322	0.5059	1.4381	3.4172	**1.7289**
5	2.2980	0.7385	**1.6970**	0.9322	1.0779	2.0100	1.6068	**1.7319**
10	3.7936	0.4430	**1.6804**	0.9322	2.1840	3.1162	0.7779	**1.6990**
20	3.9868	0.5412	**2.1575**	0.9322	4.4016	5.3338	0.3809	**1.6767**

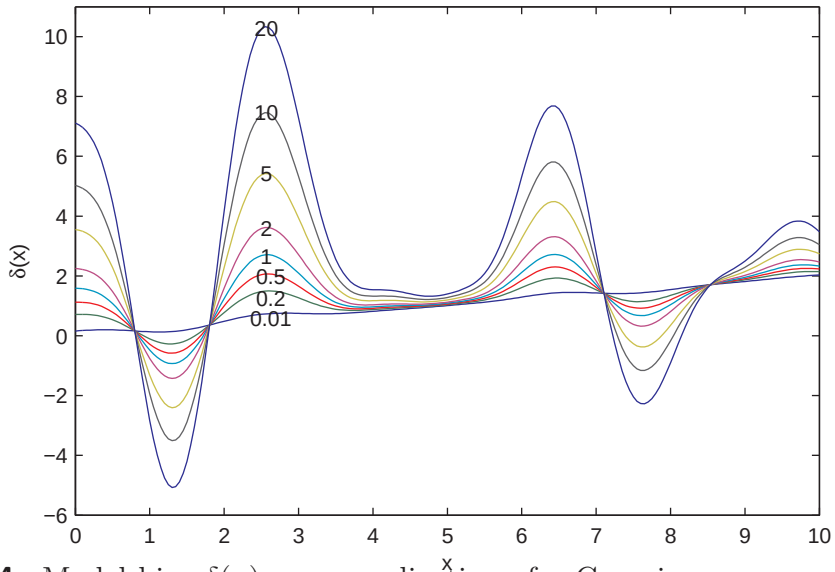


Figure 6.14: Model bias $\delta(x)$ – one realization of a Gaussian process with $\mu_\delta(x) = 0.2x$, $\phi_\delta = 1$, $P_\delta = 2$, and $\sigma^2_\delta = [0.01, 0.2, 0.5, 1, 2, 5, 10, 20]$

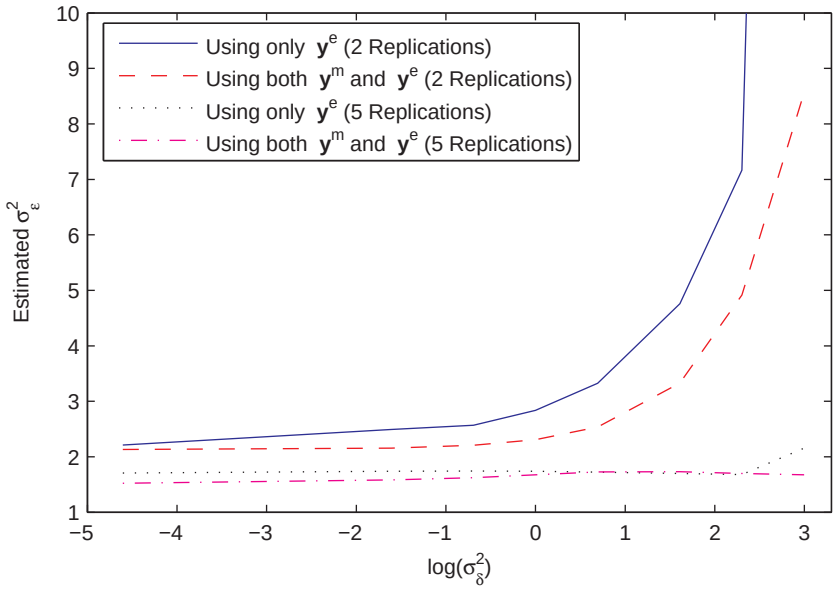


Figure 6.15: Estimated σ^2_ϵ versus $\log(\sigma^2_\delta)$

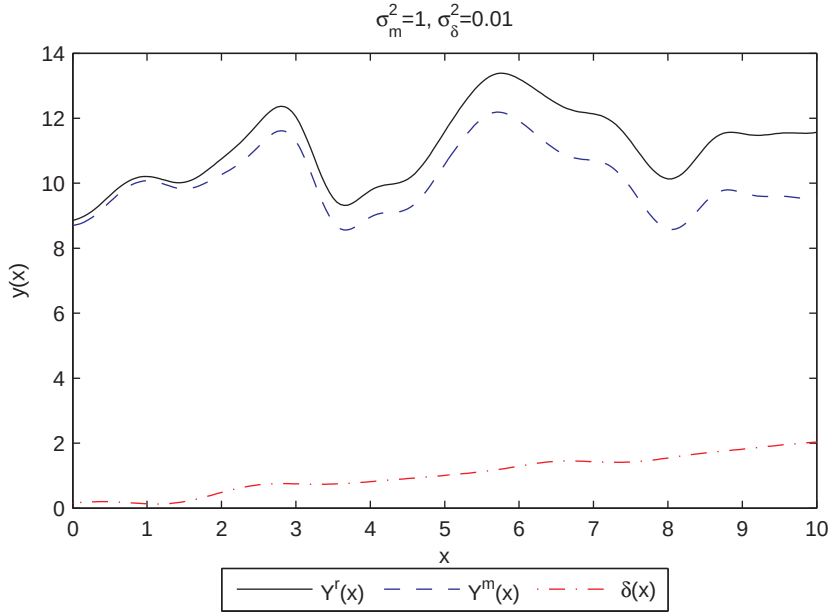


Figure 6.16: Real System Output $Y^r(x)$, Computer Model $Y^m(x)$, and Model Bias $\delta(x)$ for $\sigma_m^2 = 1$ and $\sigma_\delta^2 = 0.01$

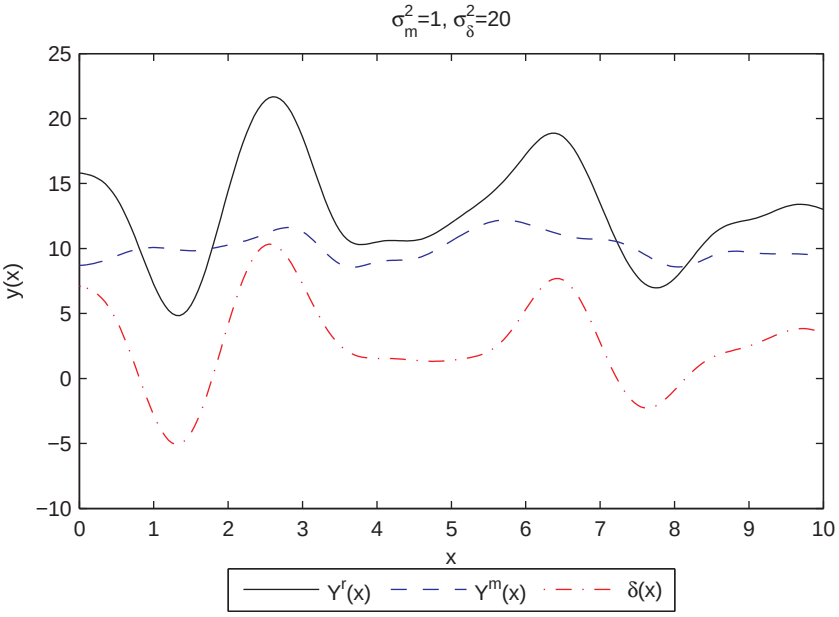


Figure 6.17: Real System Output $Y^r(x)$, Computer Model $Y^m(x)$, and Model Bias $\delta(x)$ for $\sigma_m^2 = 1$ and $\sigma_\delta^2 = 20$

6.2.3 Variance of Computer Model $Y^m(x)$, σ_m^2

Example 3: Similar to Example 2, Example 3 is different from Example 1 in only one aspect: the computer model $Y^m(x)$ is generated as one realization of a Gaussian process with $\mu_m(x) = 10$, $\phi_m = 2$, $P_m = 2$, and $\sigma_m^2 = [0.01, 0.2, 0.5, 1, 2, 5, 10, 20]$. Figure 6.18 displays computer models for different values of σ_m^2. The purpose here is to study the effects of the value of σ_m^2 on the prediction of $Y^r(x)$ and the estimation of σ_ϵ^2. We run the proposed Bayesian procedure with the number of replication as two and five. Results are shown in Tables 6.10 and 6.11 and Figure 6.19. They once again confirm the finding that, when the number of replications is larger and both computer outputs $\mathbf{y}^m$ and physical observations $\mathbf{y}^e$ are considered, we obtain a more accurate prediction of $Y^r(x)$ and estimate of σ_ϵ^2.

Tables 6.10 shows that the increase in the value of σ_m^2 causes only a slight change in RMSPEs for $Y^r(x)$ as long as $\mathbf{y}^m$ are considered (see the $\hat{Y}_m^r$ and $\hat{Y}^r$ columns). Such a result can be attributed to the fact that there are twenty points in the computer design set D_m. Therefore, we can predict the computer model $Y^m(x)$ fairly well using $\mathbf{y}^m$ no matter how large the value of σ_m^2 is (see the $\hat{Y}_m^m$ column. The RMSPEs of the predictions of $Y^m(x)$ are small). This conjecture is reinforced by the fact that the RMSPEs in the $\hat{Y}_m^r$ column are close to the square root of the mean squared bias, which equals to $\sqrt{\sum_{i=1}^{201} \delta^2(x_i)/201} = 1.6354$. On the other hand, the same increase in the value of σ_m^2 leads to a quick rise in RMSPE based on only physical observations $\mathbf{y}^e$ (see the $\hat{Y}_e^r$ column). This can be explained by the fact that, as shown in Figure 6.18, the curvature of the computer model $Y^m(x)$ increases as the value of σ_m^2 increases. As a result, the curvature of the real system output $Y^r(x)$, dominated by the model bias $\delta(x)$ when $\sigma_m^2 = 0.01$ (see Figure 6.20), become more and more dominated by the computer model $Y^m(x)$ as the value of σ_m^2 increases (see Figure 6.21). Therefore, although we can predict $Y^r(x)$ well using only $\mathbf{y}^e$ when σ_m^2 is small and the curvature of $Y^r(x)$ is dominated by the model bias $\delta(x)$, $\mathbf{y}^e$ become less capable to capture the

curvature of $Y^r(x)$ as σ_m^2 increases, and then leads to a much larger RMSPE. Both Table 6.11 and Figure 6.19 show that a larger number of replications leads to a more accurate estimate of σ_ϵ^2. Moreover, the value of σ_m^2 has no influence in the estimate of σ_ϵ^2 when both $\mathbf{y}^m$ and $\mathbf{y}^e$ are considered and little influence in the estimate of σ_ϵ^2 when only $\mathbf{y}^e$ are considered.

Comparing Table 6.10 with Table 6.8 reveals that the RMSPEs based on both $\mathbf{y}^m$ and $\mathbf{y}^e$ seems more sensitive to the value of σ_δ^2 than to the value of σ_m^2 while the RMSPEs. This might be due to the fact that there are twenty design points in D_m while only seven design points in D_e.

Table 6.10: RMSPEs of Predictions of $Y^r(x)$ or $Y^m(x)$ at 201 x values (from 0 to 10 with an increment 0.05)

Number of Replications = 2

	RMSPE			
σ_m^2	$\hat{Y}_m^r = E[Y^m\|\mathbf{y}^m]$	$\hat{Y}_e^r = E[Y^r\|\mathbf{y}^e]$	$\hat{Y}^r = E[Y^r\|\mathbf{y}^e, \mathbf{y}^m]$	$\hat{Y}_m^m = E[Y^m\|\mathbf{y}^m]$
0.01	1.6353	0.8684	0.8259	0.0160
0.2	1.6310	1.0799	0.8218	0.0435
0.5	1.6283	1.2787	0.8214	0.0732
1	1.6260	1.5185	0.8231	0.1071
2	1.6240	1.8945	0.8271	0.1542
5	1.6240	2.6941	0.8417	0.2467
10	1.6304	3.6428	0.8705	0.3502
20	1.6504	4.3510	0.9292	0.4963

Number of Replications = 5

	RMSPE			
σ_m^2	$\hat{Y}_m^r = E[Y^m\|\mathbf{y}^m]$	$\hat{Y}_e^r = E[Y^r\|\mathbf{y}^e]$	$\hat{Y}^r = E[Y^r\|\mathbf{y}^e, \mathbf{y}^m]$	$\hat{Y}_m^m = E[Y^m\|\mathbf{y}^m]$
0.01	1.6353	0.7634	0.7394	0.0160
0.2	1.6310	0.9528	0.7307	0.0435
0.5	1.6283	1.1211	0.7243	0.0732
1	1.6260	1.3736	0.7293	0.1071
2	1.6240	1.7041	0.7335	0.1542
5	1.6240	2.3145	0.7458	0.2467
10	1.6304	3.0627	0.7806	0.3502
20	1.6504	4.2062	0.8361	0.4963

Table 6.11: Estimated Experimental Error Variance σ_ϵ^2

Number of Replications = 2

	Using only $\mathbf{y}^e$			Using both $\mathbf{y}^e$ and $\mathbf{y}^m$				
σ_m^2	$\hat{\sigma}_r^2$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_r^2)$	$\hat{\sigma}_m^2$	$\hat{\sigma}_\delta^2$	$\hat{\sigma}_r^2$ $(\hat{\sigma}_m^2 + \hat{\sigma}_\delta^2)$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_\delta^2)$
0.01	0.3469	6.6529	**2.3076**	0.4283	0.3468	0.7752	6.6522	**2.3072**
0.2	0.3525	6.7974	**2.3962**	0.3185	0.3468	0.6653	6.6522	**2.3072**
0.5	0.3616	7.0646	**2.5545**	0.5404	0.3468	0.8872	6.6522	**2.3072**
1	0.3748	7.5679	**2.8364**	0.9322	0.3468	1.2790	6.6522	**2.3072**
2	0.3935	8.7688	**3.4503**	1.7233	0.3468	2.0701	6.6522	**2.3072**
5	0.4094	13.6358	**5.5825**	4.1024	0.3468	4.4492	6.6522	**2.3072**
10	0.4029	23.7101	**9.5517**	8.0695	0.3468	8.4163	6.6522	**2.3072**
20	6.6490	0.5494	**3.6530**	16.0045	0.3468	16.3513	6.6522	**2.3072**

Number of Replications = 5

	Using only $\mathbf{y}^e$			Using both $\mathbf{y}^e$ and $\mathbf{y}^m$				
σ_m^2	$\hat{\sigma}_r^2$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_r^2)$	$\hat{\sigma}_m^2$	$\hat{\sigma}_\delta^2$	$\hat{\sigma}_r^2$ $(\hat{\sigma}_m^2 + \hat{\sigma}_\delta^2)$	$\hat{\tau}$	$\hat{\sigma}_\epsilon^2$ $(\hat{\tau} \cdot \hat{\sigma}_\delta^2)$
0.01	0.3941	4.2754	**1.6850**	0.4283	0.3772	0.8056	4.4357	**1.6733**
0.2	0.4987	3.4636	**1.7273**	0.3185	0.3772	0.6957	4.4357	**1.6733**
0.5	0.6457	2.7002	**1.7435**	0.5404	0.3772	0.9176	4.4357	**1.6733**
1	0.9059	1.9197	**1.7391**	0.9322	0.3772	1.3094	4.4357	**1.6733**
2	1.4413	1.1918	**1.7179**	1.7233	0.3772	2.1005	4.4357	**1.6733**
5	2.9949	0.5634	**1.6875**	4.1024	0.3772	4.4796	4.4357	**1.6733**
10	5.4633	0.3060	**1.6720**	8.0695	0.3772	8.4467	4.4357	**1.6733**
20	10.2244	0.1626	**1.6627**	16.0045	0.3772	16.3817	4.4357	**1.6733**

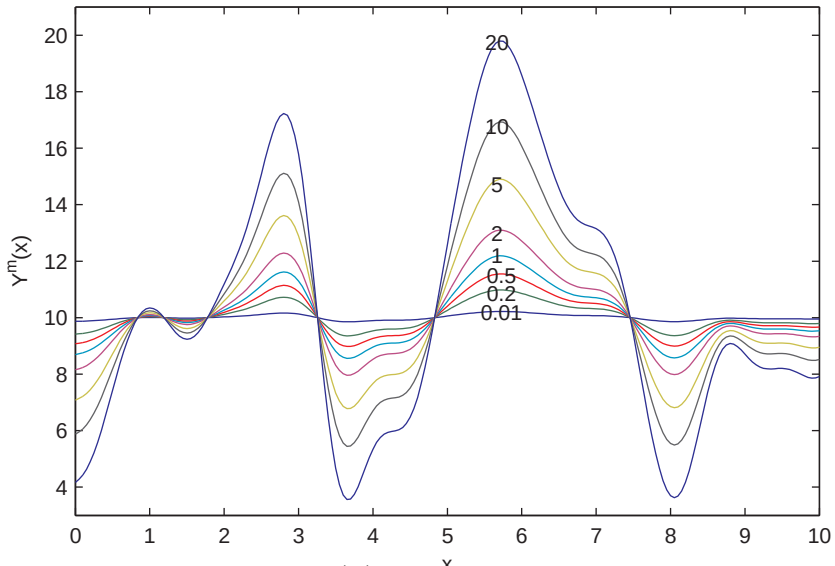


Figure 6.18: Computer Model $Y^m(x)$ – one realization of a Gaussian process with $\mu_m(x) = 10$, $\phi_m = 2$, $P_m = 2$, and $\sigma_m^2 = [0.01, 0.2, 0.5, 1, 2, 5, 10, 20]$

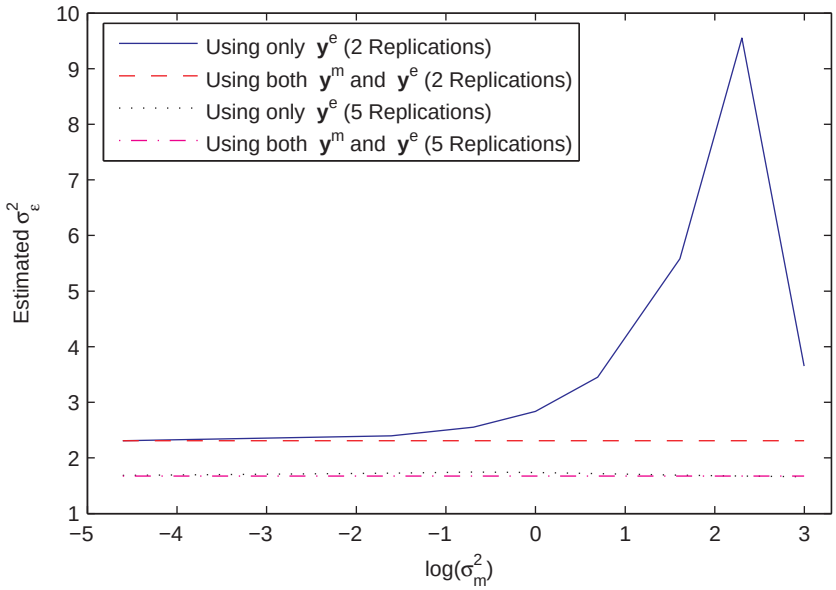


Figure 6.19: Estimated σ_ϵ^2 versus $\log(\sigma_m^2)$

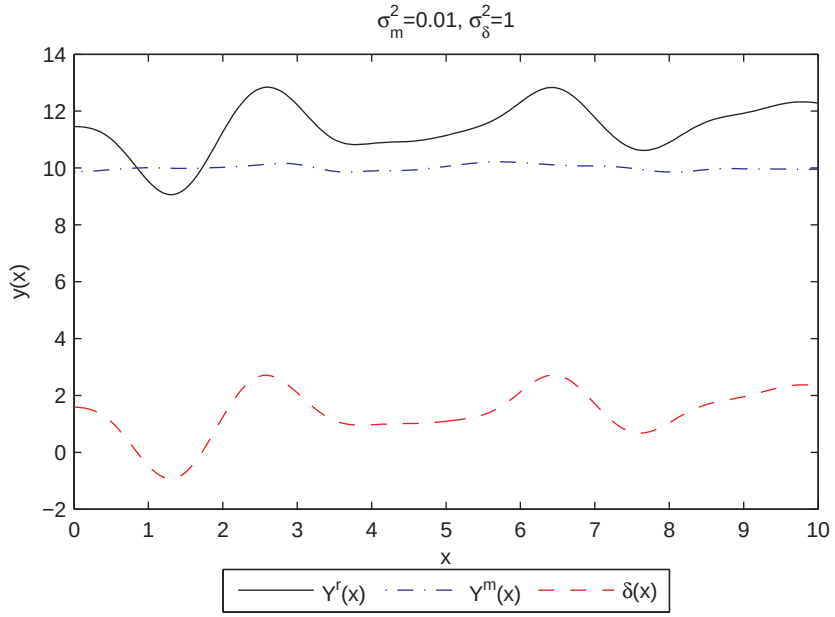


Figure 6.20: Real System Output $Y^r(x)$, Computer Model $Y^m(x)$, and Model Bias $\delta(x)$ for $\sigma_m^2 = 0.01$ and $\sigma_\delta^2 = 1$

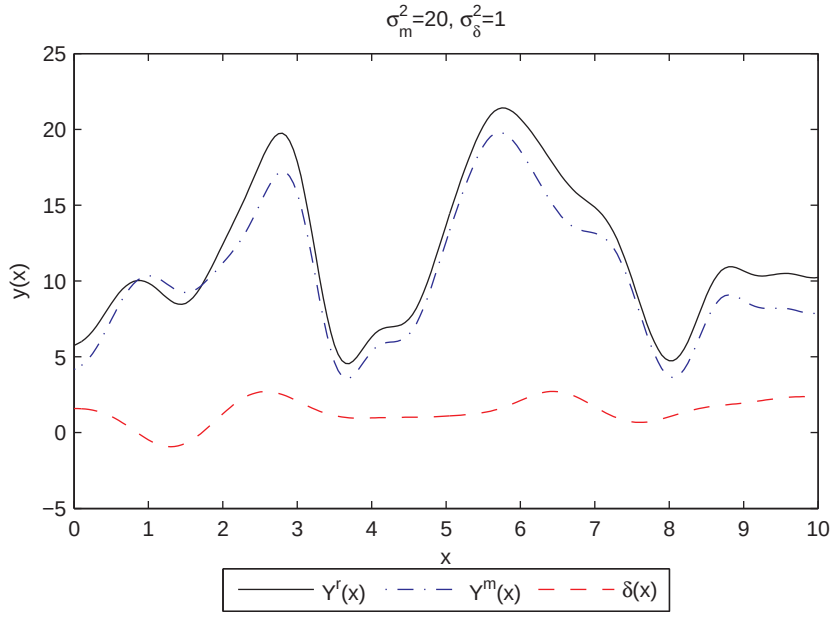


Figure 6.21: Real System Output $Y^r(x)$, Computer Model $Y^m(x)$, and Model Bias $\delta(x)$ for $\sigma_m^2 = 20$ and $\sigma_\delta^2 = 1$

6.2.4 Conclusions

We have used three simulated numerical examples to investigate the performance of the proposed Bayesian approach and study the influences of three factors, the number of replications in physical experiments, the variance of $\delta(x)$, and the variance of $Y^m(x)$. From the results obtained in the three examples, we can draw the following conclusions on the proposed approach:

- The choices of design sets D_m and D_e are crucial to the performance of the proposed approach.

  A design set is determined by its size (i.e., the number of points in the design set) and location (i.e., the locations of points in the design set). Since computer outputs are usually less expensive and time-consuming compared to physical observations, we can have a relatively large computer design set D_m. For example, in all three examples, we have twenty points in the computer design set D_m while only seven points in the physical design set D_e. The locations of design points should be chosen such that the major characteristics of the computer model $Y^m(x)$ and the real system output $Y^r(x)$ are captured. For example, in Example 1, physical observations in the region of $x > 3$ capture the maxima and minima of $Y^r(x)$. As a result, the predictions of $Y^r(x)$ in this region are closer to the corresponding true values.

- As the number of replications increases, the proposed approach performs better. All three examples show that a larger number of replications leads to a more accurate and stable prediction of $Y^r(x)$ and estimate of σ_ϵ^2.

- The values of σ_δ^2 and σ_m^2 affect the behavior of $Y^r(x)$ and therefore affect the performance of the proposed approach.

  When $\sigma_m^2 \gg \sigma_\delta^2$, the curvature of $Y^r(x)$ is dominated by the computer model $Y^m(x)$ (see Figures 6.16 and 6.21). In other words, when $\sigma_m^2 \gg \sigma_\delta^2$, the computer

model $Y^m(x)$ might not be an accurate representation of the real system $Y^r(x)$, but it captures the shape of $Y^r(x)$. When $\sigma_m^2 \ll \sigma_\delta^2$, the curvature of $Y^r(x)$ is dominated by the model bias $\delta(x)$ (see Figures 6.17 and 6.20), and the computer model $Y^m(x)$ does not capture the behavior of $Y^r(x)$. As a result, since we often have a large D_m and a small D_e (therefore $Y^m(x)$ can be predicted better than $\delta(x)$), the proposed approach performs better for smaller σ_δ^2 ($\delta(x)$ with smaller σ_δ^2 is easier to predict) and the value of σ_m^2 has little influence on the performance. Furthermore, with the increase of $\sigma_m^2/\sigma_\delta^2$ (i.e., the curvature of $Y^r(\mathbf{x})$ becomes more dominated by the computer model $Y^m(\mathbf{x})$), using both $\mathbf{y}^m$ and $\mathbf{y}^e$ leads to a larger improvement in prediction accuracy compared to using only $\mathbf{y}^e$.

6.3 A Generalization to The Proposed Bayesian Approach

The proposed approach assumes that the two Gaussian processes, the computer model $Y^m(\mathbf{x})$ and the model bias $\delta(\mathbf{x})$, are mutually independent. Such independence assumption may not hold in reality. For example, it is possible that the model bias $\delta(\mathbf{x})$ is positively correlated with the computer output $Y^m(\mathbf{x})$. That is, $\delta(x)$ is large where $Y^m(\mathbf{x})$ is large and small where $Y^m(\mathbf{x})$ is small. In this section, we derive the posterior distributions of $\delta(\mathbf{x})$ and $Y^r(\mathbf{x})$ given computer outputs $\mathbf{y}^m$ and physical observations $\mathbf{y}^e$ when $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ in equation (6.1) are assumed to be correlated.

6.3.1 Correlation between $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$

Instead of assuming that $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ are mutually independent, we assume that

- $(Y^m(\mathbf{x}), \delta(\mathbf{x}))^T$ has a bivariate normal distribution

$$\begin{bmatrix} Y^m(\mathbf{x}) \\ \delta(\mathbf{x}) \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{f}_m^T(\mathbf{x})\boldsymbol{\beta}_m \\ \mathbf{f}_\delta^T(\mathbf{x})\boldsymbol{\beta}_\delta \end{bmatrix}, \begin{bmatrix} \sigma_m^2 & \rho\sigma_m\sigma_\delta \\ \rho\sigma_m\sigma_\delta & \sigma_\delta^2 \end{bmatrix} \right) \tag{6.2}$$

where ρ obviously is the correlation between $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$, taking values in the interval $[-1, 1]$.

- $\text{Cov}(Y^m(\mathbf{x}_i), \delta(\mathbf{x}_j)) = 0$ for any $i \neq j$.
- $p(\boldsymbol{\theta}, \rho) = p(\boldsymbol{\theta}) \cdot p(\rho)$, where $\boldsymbol{\theta} = \{\boldsymbol{\beta}_m, \sigma_m^2, \boldsymbol{\phi}_m, \mathbf{P}_m, \boldsymbol{\beta}_\delta, \sigma_\delta^2, \boldsymbol{\phi}_\delta, \mathbf{P}_\delta, \sigma_\epsilon^2\}$, the collection of all parameters except the new correlation parameter ρ in equation (6.1).

All the other assumptions made in Chapter 5 including the prior distributions for $\boldsymbol{\beta}_m$, σ_m^2, $\boldsymbol{\beta}_\delta$, and σ_δ^2 remain the same. As a result, when the correlation parameter ρ equals to zero, $Y^m(\mathbf{x})$ and $\delta(\mathbf{x})$ are independent, and we have exactly the same model as before.

6.3.2 Posterior Distributions of $\delta(\mathbf{x})$ and $Y^m(\mathbf{x})$

In this subsection, we derive the posterior distributions of $\delta(\mathbf{x})$ and $Y^m(\mathbf{x})$ when $D_e \subseteq D_m$. Without loss of generality, we rearrange the vector of computer outputs $\mathbf{y}^m$ such as the first n_e elements are computer outputs at $D_e = \{\mathbf{x}_1^e, \cdots, \mathbf{x}_{n_e}^e\}$, denoted by $\mathbf{y}_{n_e}^m = (y^m(\mathbf{x}_1^e), \cdots, y^m(\mathbf{x}_{n_e}^e))^T$.

Given $\boldsymbol{\theta}$ and ρ, for any set $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ in the input space, we derive the posteriors of $\delta(D)$ and $Y^m(D)$ given $\mathbf{y}^m$ and $\mathbf{y}^e$. For $\delta(D)$, we consider two cases:

- $D \cap D_m = \emptyset$.
- $D = D_m - D_e$, where $D_m - D_e$ is the complement of D_e in D_m.

For $Y^m(D)$, we need consider only the first case since, in the second case, the values of $Y^m(D)$ are available.

• Distribution of $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}, \rho$ with $D \cap D_m = \emptyset$

According to equation (6.1), given $\boldsymbol{\theta}$ and ρ, for any set $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ in the input

space such that $D \cap D_m = \emptyset$, we have that

$$\begin{bmatrix} \delta(D) \\ \mathbf{y}^e \\ \mathbf{y}^m \end{bmatrix} \Bigg| \boldsymbol{\theta}, \rho \sim N(\begin{bmatrix} \mathbf{F}_\delta(D)\boldsymbol{\beta}_\delta \\ \mathbf{F}_m(D_e)\boldsymbol{\beta}_m + \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta \\ \mathbf{F}_m(D_m)\boldsymbol{\beta}_m \end{bmatrix}, \tag{6.3}$$
$$\begin{bmatrix} \sigma_\delta^2\mathbf{R}_\delta(D) & \sigma_\delta^2\mathbf{R}_\delta(D, D_e) & \mathbf{0}_{n\times n_m} \\ \sigma_\delta^2\mathbf{R}_\delta(D_e, D) & \Sigma_e & \Sigma_{em} \\ \mathbf{0}_{n_m\times n} & \Sigma_{em}^T & \sigma_m^2\mathbf{R}_m(D_m) \end{bmatrix})$$

where

$$\Sigma_e = \mathrm{Cov}(\mathbf{y}^e, \mathbf{y}^e) = \sigma_m^2\mathbf{R}_m(D_e) + \sigma_\delta^2\mathbf{R}_\delta(D_e) + 2\rho\sigma_m\sigma_\delta\mathbf{I}_{n_e} + \sigma_\epsilon^2\mathbf{I}_{n_e} \tag{6.4a}$$

$$\Sigma_{em} = \mathrm{Cov}(\mathbf{y}^e, \mathbf{y}^m) = \sigma_m^2\mathbf{R}_m(D_e, D_m) + \rho\sigma_m\sigma_\delta[\ \mathbf{I}_{n_e} \quad \mathbf{0}_*\ ] \tag{6.4b}$$

$$\mathbf{0}_* = \mathbf{0}_{n_e\times(n_m-n_e)} \tag{6.4c}$$

and $\mathbf{0}_{k_1\times k_2}$ is a $k_1 \times k_2$ matrix of zeros. Therefore, $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}, \rho$ has a multivariate normal distribution with mean vector

$$\mathbf{F}_\delta(D)\boldsymbol{\beta}_\delta + \begin{bmatrix} \sigma_\delta^2\mathbf{R}_\delta(D, D_e) & \mathbf{0}_{n\times n_m} \end{bmatrix} \Sigma^{-1} \begin{bmatrix} \mathbf{y}^e - \mathbf{F}_m(D_e)\boldsymbol{\beta}_m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta \\ \mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m \end{bmatrix} \tag{6.5}$$

and covariance matrix

$$\sigma_\delta^2\mathbf{R}_\delta(D) - \begin{bmatrix} \sigma_\delta^2\mathbf{R}_\delta(D, D_e) & \mathbf{0}_{n\times n_m} \end{bmatrix} \Sigma^{-1} \begin{bmatrix} \sigma_\delta^2\mathbf{R}_\delta(D_e, D) \\ \mathbf{0}_{n_m\times n} \end{bmatrix}. \tag{6.6}$$

where

$$\Sigma = \begin{bmatrix} \Sigma_e & \Sigma_{em} \\ \Sigma_{em}^T & \sigma_m^2\mathbf{R}_m(D_m) \end{bmatrix} \tag{6.7}$$

After some matrix manipulations, we have that $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}, \rho$ has a multivariate normal distribution with mean vector

$$\mathbf{F}_\delta(D)\boldsymbol{\beta}_\delta + \mathbf{R}_\delta(D, D_e)\mathbf{Q}^{-1}\left[\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta - \rho\frac{\sigma_\delta}{\sigma_m}\mathbf{P}(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m)\right] \tag{6.8}$$

and covariance matrix

$$\sigma_\delta^2 \left[\mathbf{R}_\delta(D) - \mathbf{R}_\delta(D, D_e)\mathbf{Q}^{-1}\mathbf{R}_\delta(D_e, D)\right], \tag{6.9}$$

where

$$\mathbf{P} = [\mathbf{I}_{n_e} \;\; \mathbf{0}_*]\, \mathbf{R}_m^{-1}(D_m) \tag{6.10a}$$

$$\mathbf{Q} = \mathbf{R}_\delta(D_e) + \tau \mathbf{I}_{n_e} - \rho^2 [\mathbf{I}_{n_e} \;\; \mathbf{0}_*] \cdot \mathbf{R}_m^{-1}(D_m) \cdot \begin{bmatrix} \mathbf{I}_{n_e} \\ \mathbf{0}_*^T \end{bmatrix}. \tag{6.10b}$$

• **Distribution of $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}, \rho$ with $D = D_m - D_e$**

For $D = D_m - D_e$, the distribution of

$$\begin{bmatrix} \delta(D) \\ \mathbf{y}^e \\ \mathbf{y}^m \end{bmatrix} \Bigg| \boldsymbol{\theta}, \rho \tag{6.11}$$

is the same as that shown in equation (6.3) except that the sub-matrix $\mathbf{0}_{n \times n_m}$ in the covariance matrix is replaced by

$$\rho \sigma_m \sigma_\delta \begin{bmatrix} \mathbf{0}_*^T & \mathbf{I}_{n_m - n_e} \end{bmatrix}. \tag{6.12}$$

Therefore, when $D = D_m - D_e$, $\delta(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}, \rho$ has a multivariate normal distribution with mean vector

$$\begin{aligned} &\mathbf{F}_\delta(D)\boldsymbol{\beta}_\delta + \mathbf{R}_\delta(D, D_e)\mathbf{Q}^{-1} \left[\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta - \rho \frac{\sigma_\delta}{\sigma_m} \mathbf{P}(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m)\right] \\ &+ \left[\mathbf{0}_*^T \;\; \mathbf{I}_{n_m - n_e}\right] \left[\rho \frac{\sigma_\delta}{\sigma_m} \mathbf{Q}_1^{-1}(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m) - \rho^2 \mathbf{P}^T \mathbf{Q}^{-1}(\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta)\right] \end{aligned} \tag{6.13}$$

and covariance matrix

$$\begin{aligned} &\sigma_\delta^2 \big[\mathbf{R}_\delta(D) - \rho^2 \; [\mathbf{0}_* \;\; \mathbf{I}_{n_m - n_e}]\, \mathbf{R}_m^{-1}(D_m) \begin{bmatrix} \mathbf{0}_* \\ \mathbf{I}_{n_m - n_e} \end{bmatrix} \\ &-(\mathbf{R}_\delta(D_e, D) - \rho^2 \, \mathbf{P} \begin{bmatrix} \mathbf{0}_* \\ \mathbf{I}_{n_m - n_e} \end{bmatrix})^T \mathbf{Q}^{-1} \left(\mathbf{R}_\delta(D_e, D) - \rho^2 \, \mathbf{P} \begin{bmatrix} \mathbf{0}_* \\ \mathbf{I}_{n_m - n_e} \end{bmatrix}\right)\big] \end{aligned} \tag{6.14}$$

• **Distribution of $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}, \rho$**

According to equation (6.1), given $\boldsymbol{\theta}$ and ρ, for any set $D = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ in the input space such that $D \cap D_m = \emptyset$, we have that

$$\begin{bmatrix} Y^m(D) \\ \mathbf{y}^e \\ \mathbf{y}^m \end{bmatrix} \Bigg| \boldsymbol{\theta}, \rho \sim N\Bigg(\begin{bmatrix} \mathbf{F}_m(D)\boldsymbol{\beta}_m \\ \mathbf{F}_m(D_e)\boldsymbol{\beta}_m + \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta \\ \mathbf{F}_m(D_m)\boldsymbol{\beta}_m \end{bmatrix}, \tag{6.15}$$
$$\begin{bmatrix} \sigma_m^2\mathbf{R}_m(D) & \sigma_m^2\mathbf{R}_m(D, D_e) & \sigma_m^2\mathbf{R}_m(D, D_m) \\ \sigma_m^2\mathbf{R}_m(D_e, D) & \Sigma_e & \Sigma_{em} \\ \sigma_m^2\mathbf{R}_m(D_m, D) & \Sigma_{em}^T & \sigma_m^2\mathbf{R}_m(D_m) \end{bmatrix} \Bigg).$$

This gives the distribution of $Y^m(D)|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}, \rho$ as a multivariate normal distribution with mean vector

$$\mathbf{F}_m(D)\boldsymbol{\beta}_m + \mathbf{R}_m(D, D_m)\left[\mathbf{Q}_1^{-1}(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m) - \rho\frac{\sigma_m}{\sigma_\delta}\mathbf{P}^T\mathbf{Q}^{-1}(\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta)\right] \tag{6.16}$$

and covariance matrix

$$\sigma_m^2\left[\mathbf{R}_m(D) - \mathbf{R}_m(D, D_m)\mathbf{Q}_1^{-1}\mathbf{R}_m(D_m, D)\right], \tag{6.17}$$

where

$$\mathbf{Q}_1 = \mathbf{R}_m(D_m) - \rho^2 \begin{bmatrix} \mathbf{I}_{n_e} \\ \mathbf{0}_* \end{bmatrix} (\mathbf{R}_\delta(D_e) + \tau\mathbf{I}_{n_e})^{-1} [\mathbf{I}_{n_e}\ \mathbf{0}_*] \tag{6.18}$$

6.3.3 Full Conditional Distributions of $\boldsymbol{\beta}_m$, σ_m^2, $\boldsymbol{\beta}_\delta$, and σ_δ^2

The posterior distributions of $\delta(D)$ and $Y^m(D)$ derived in subsection 6.3.2 are conditional on parameters $\boldsymbol{\theta}$ and ρ. Instead of integrating out $\boldsymbol{\beta}_\delta$, σ_δ^2, $\boldsymbol{\beta}_m$, and σ_m^2 as we did Chapter 5, we derive the full conditional distributions of $\boldsymbol{\beta}_\delta$, σ_δ^2, $\boldsymbol{\beta}_m$, and σ_m^2 given $\mathbf{y}^e$ and $\mathbf{y}^m$ so that a Markov Chain Monte Carlo (MCMC) algorithm, Gibbs sampling, can be used to estimate their values. Those estimated values are then plugged into

equations (6.8), (6.9), (6.13), (6.14), (6.14), (6.16), and (6.17) to get the posterior distributions of $\delta(D)$ and $Y^m(D)$.

• **Full Conditional Distribution of $\boldsymbol{\beta}_m$**

The full conditional distribution $\boldsymbol{\beta}_m$ given $\mathbf{y}^e$ and $\mathbf{y}^m$ can be derived by using the fact that

$$p(\boldsymbol{\beta}_m|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\boldsymbol{\beta}_m)}, \rho) \ \propto \ p(\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\theta}, \rho) \cdot p(\mathbf{y}^m|\boldsymbol{\theta}, \rho) \cdot p(\boldsymbol{\beta}_m|\sigma_m^2), \tag{6.19}$$

where

$$\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\theta}, \rho \sim N(\mathbf{y}_{n_e}^m + \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta + \rho\frac{\sigma_\delta}{\sigma_m} \cdot \mathbf{P}(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m), \ \sigma_\delta^2\mathbf{Q}). \tag{6.20}$$

After some matrix manipulations, we have that the full conditional distribution $\boldsymbol{\beta}_m$ given $\mathbf{y}^e$ and $\mathbf{y}^m$ is a multivariate normal distribution

$$\boldsymbol{\beta}_m|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\boldsymbol{\beta}_m)}, \rho \ \sim \ N(\mathbf{A}_m\mathbf{v}_m, \ \sigma_m^2\mathbf{A}_m) \tag{6.21}$$

where

$$\mathbf{A}_m^{-1} = \mathbf{F}_m^T(D_m)\mathbf{Q}_1^{-1}\mathbf{F}_m(D_m) + \mathbf{V}_m^{-1}, \tag{6.22a}$$

$$\mathbf{v}_m = \mathbf{F}_m^T(D_m)\left[\mathbf{Q}_1^{-1}\mathbf{y}^m - \rho\frac{\sigma_m}{\sigma_\delta}\mathbf{P}^T\mathbf{Q}^{-1}(\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta((D_e)\boldsymbol{\beta}_\delta)\right] + \mathbf{V}_m^{-1}\mathbf{b}_m \tag{6.22b}$$

and $\boldsymbol{\theta}_{-(\cdot)}$ contains all the parameters except those inside the parentheses.

• **Full Conditional Distribution of σ_m^2**

Similarly, we derive the full conditional distribution of σ_m^2 given $\mathbf{y}^e$ and $\mathbf{y}^m$ using the fact that

$$p(\sigma_m^2|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\sigma_m^2)}, \rho) \ \propto \ p(\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\theta}, \rho) \cdot p(\mathbf{y}^m|\boldsymbol{\theta}, \rho) \cdot p(\boldsymbol{\beta}_m|\sigma_m^2) \cdot p(\sigma_m^2) \tag{6.23}$$

which gives

$$p(\sigma_m^2|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\sigma_m^2)}, \rho) = C_0^m \cdot (\sigma_m^2)^{-\frac{q_m}{2}-\frac{n_m}{2}-\alpha_m-1} \cdot \exp\left\{-\frac{C_1^m}{\sigma_m^2} + \frac{C_2^m}{\sigma_m}\right\}, \tag{6.24}$$

where

$$C_1^m = \gamma_m + \frac{1}{2}\big[(\boldsymbol{\beta}_m - \mathbf{b}_m)^T \mathbf{V}_m^{-1}(\boldsymbol{\beta}_m - \mathbf{b}_m) \tag{6.25a}$$
$$+(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m)^T \mathbf{Q}_1^{-1}(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m)\big],$$
$$C_2^m = \rho \cdot \frac{1}{\sigma_\delta} \cdot (\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta)^T \mathbf{Q}^{-1}\mathbf{P}(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m), \tag{6.25b}$$

and C_0^m is a normalization constant such that the term to the right side of the = sign in equation (6.24) is a density.

• Full Conditional Distribution of $\boldsymbol{\beta}_\delta$

We derive the full conditional distribution $\boldsymbol{\beta}_\delta$ given $\mathbf{y}^e$ and $\mathbf{y}^m$ using the fact that

$$p(\boldsymbol{\beta}_\delta|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\boldsymbol{\beta}_\delta)}, \rho) \;\propto\; p(\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\theta}, \rho) \cdot p(\boldsymbol{\beta}_\delta|\sigma_\delta^2), \tag{6.26}$$

which gives

$$\boldsymbol{\beta}_\delta|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\boldsymbol{\beta}_\delta)}, \rho \;\sim\; N(\mathbf{A}_\delta \mathbf{v}_\delta,\; \sigma_\delta^2 \mathbf{A}_\delta), \tag{6.27}$$

where

$$\mathbf{A}_\delta^{-1} = \mathbf{F}_\delta^T(D_e)\mathbf{Q}^{-1}\mathbf{F}_\delta(D_e) + \mathbf{V}_\delta^{-1}, \tag{6.28a}$$
$$\mathbf{v}_\delta = \mathbf{F}_\delta^T(D_e)\mathbf{Q}^{-1}\left[\mathbf{y}^e - \mathbf{y}_{n_e}^m - \rho\frac{\sigma_\delta}{\sigma_m}\mathbf{P}(\mathbf{y}^m - \mathbf{F}_m((D_m)\boldsymbol{\beta}_m)\right] + \mathbf{V}_\delta^{-1}\mathbf{b}_\delta \tag{6.28b}$$

• Full Conditional Distribution of σ_δ^2

We derive the full conditional distribution σ_δ^2 given $\mathbf{y}^e$ and $\mathbf{y}^m$ using the fact that

$$p(\sigma_\delta^2|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\sigma_\delta^2)}, \rho) \;\propto\; p(\mathbf{y}^e|\mathbf{y}^m, \boldsymbol{\theta}, \rho) \cdot p(\boldsymbol{\beta}_\delta|\sigma_\delta^2) \cdot p(\sigma_\delta^2) \tag{6.29}$$

which gives

$$p(\sigma_\delta^2|\mathbf{y}^e, \mathbf{y}^m, \boldsymbol{\theta}_{-(\sigma_\delta^2)}, \rho) = C_0^\delta \cdot (\sigma_\delta^2)^{-\frac{q_\delta}{2}-\frac{n_e}{2}-\alpha_\delta-1} \cdot \exp\left\{-\frac{C_1^\delta}{\sigma_\delta^2} + \frac{C_2^\delta}{\sigma_\delta}\right\} \tag{6.30}$$

where

$$C_1^\delta = \gamma_\delta + \frac{1}{2}\big[(\boldsymbol{\beta}_\delta - \mathbf{b}_\delta)^T \mathbf{V}_\delta^{-1}(\boldsymbol{\beta}_\delta - \mathbf{b}_\delta) \tag{6.31a}$$
$$+(\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta)^T \mathbf{Q}^{-1}(\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta)\big],$$
$$C_2^\delta = \rho \cdot \frac{1}{\sigma_m} \cdot (\mathbf{y}^e - \mathbf{y}_{n_e}^m - \mathbf{F}_\delta(D_e)\boldsymbol{\beta}_\delta)^T \mathbf{Q}^{-1}\mathbf{P}(\mathbf{y}^m - \mathbf{F}_m(D_m)\boldsymbol{\beta}_m), \tag{6.31b}$$

and C_0^δ is a normalization constant such that the term to the right side of the = sign in equation (6.30) is a density.

CHAPTER VII

SUMMARY AND FUTURE RESEARCH

7.1 Exponential Smoothing for Forecasting

7.1.1 Summary

In this research, we investigated three types of statistical models that have been found to underlie ES methods. They are ARIMA model, MSOE state space model, and SSOE state space model. We established the relationship among the three classes of statistical models and concluded that the class of SSOE state space models is broader than the other two and provides a general statistical framework for the study of ES methods. To better understand ES methods, we investigated the performance of ES methods on time series of ARIMA-type.

We then continued to propose a new forecasting method, ESCov. This new method incorporates covariates into ES methods and uses ES methods to model what left unexplained in the time series of interest by covariates. Numerical studies based on two real-life examples demonstrated that ESCov outperforms ES methods and regression models with ARIMA errors. We identified underlying SSOE state space models for ESCov, discussed ML estimation using underlying SSOE models, and derived the variances of forecasts by ESCov for the construction of prediction intervals. We also suggested a two-step model selection procedure to choose covariates and ES methods in the use of ESCov.

7.1.2 Future Research

We have tested ESCov on two real-life examples. More work is needed to explore the performance of ESCov and to understand when ESCov performs well and why. We

have assumed the time series being forecast and covariates have a linear relationship with fixed coefficients. Future work would be to generalize ESCov to handle time-varying coefficients and nonlinear relationships.

7.2 Bayesian Validation of Computer Models

7.2.1 Summary

In this research, we proposed a Bayesian approach to the validation of computer models. This approach integrates computer outputs and physical observations together to give an accurate prediction of the output of the real system for which the computer model is built. The prediction of the real system output is then used to validate the computer model. The performance of the proposed approach was tested on real-life examples and investigated through the use of simulated examples. We also proposed a generalization to the proposed approach.

7.2.2 Future Research

We have investigated the impacts of three factors (the number of replications in physical experiments, the variance of $\delta(\mathbf{x})$, and the variance of $Y^m(\mathbf{x})$) on the performance of the proposed Bayesian approach. More studies are needed to explore the impacts of those three factors and other factors, such as design sets D_e and D_m, correlation parameters $\boldsymbol{\phi}_\delta$ and $\boldsymbol{\phi}_m$, and prior distribution parameters. We have proposed a generalization to the proposed approach by assuming that $Y^m(\mathbf{x})$ and $\delta(x)$ are correlated. The next step along this direction would be the implementation of the generalized approach and the investigation of its performance, such as the impacts of the correlation parameter.

REFERENCES

[1] American Institute of Aeronautics and Astronautics (1998). *Guide for the verification and validation of computational fluid daynamics simulations*. AIAA-G-077-1998.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second international symposium on information theory*, edited by Petrov, B. N. and Csaki, F., Akademiai Kiado, Budapest.

[3] Bayarri, M. J., Berger, J. O., Higdon, D., Kennedy, M. C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H., and Tu, J. (2002). A Framework for Validation of Computer Models. *Foundations for Verification and Validation in the 21st Century Workshop*, Johns Hopkins University.

[4] Billah, B., King, M. L., Snyder, R. D., Koehler, A. B. (2006). Exponential smoothing model selection for forecasting. *International Journal of Forecasting*, 22(2), 239-247

[5] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. second edition, Springer-Verlag, New York.

[6] Brown, R. G. (1959). *Statistical Forecasting for Inventory Control*. McGraw-Hill

[7] Brown, R. G. (1963). *Smoothing, Forecasting and Prediction of Discrete Time series*. Prentice-Hall

[8] Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time series analysis: forecasting and control*. 3rd ed, Prentice-Hall, Inc.

[9] Castillo, E. D. (2001). Some properties of EWMA feedback quality adjustment schemes for drifting distribution. *Journal of Quality Control*, 33, 153-166.

[10] Chatfield, C. (1996). *The Analysis of Time Series: An Introduction*. fifth edition, Chapman and Hall Ltd, London.

[11] Chatfield, C., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2001). A new look at models for exponential smoothing. *The Statistican*, 50, 147-159.

[12] Chen, C. (1997). Robustness properties of some forecasting methods for seasonal time series: a Monte Carlo study. *International Journal of Forecasting*, 13, 269-280.

[13] Cohen, G. D. (1963). A note on exponential smoothing and autocorrelated inputs. *Operations Research*, 2, 361-367.

[14] Cogger, K. O. (1973). Specification analysis. *Journal of the American Statistical Association*, 68, 899-905.

[15] Cox, D. R. (1961). Prediction by exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society (B)*, 23, 414-422.

[16] Dewettinck, K., Visscher A. D., Deroo, L., and Huyghebaert, A. (1999). Modeling the steady-state thermodynamic operation point of top-spray fluidized bed processing. *Journal of Food Engineering*, 39, 131-143

[17] Duncan, D. B. and Horn, S. D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis. *Journal of the American Statistical Association*, 67, 815-821.

[18] Easterling, R. G. and Berger, J. O. (2002). Statistical Foundations for The Validation of Computer Models. *Foundations for Verification and Validation in the 21st Century Workshop*, Johns Hopkins University

[19] Fuller, W. A. (1996). *Introduction to statistical time series*. 2nd edition. John Wiley and Sons, New York.

[20] Gardner, E. Jr.(1985). Exponential smoothing: the state of the art. *Journal of Forecasting*, 4(1), 1-28.

[21] Gardner, E. Jr. and Mckenzie, E. (1985). Forecasting trends in time series. *Management Science*, 31, 1237-1246.

[22] Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society (B)*, 41, 190-195.

[23] Harrison, P. J. (1967). Exponential smoothing and short-time sales forecasting. *Management Science*, 13, 821-842.

[24] Harvey, A. C. (1984). A unified view of statistical forecasting procedures. *Journal of Forecasting*, 3(3), 245-275.

[25] Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and Kalman Filter*, Cambridge University Press, Cambridge.

[26] Harvey, A. C. (1993). *Time Series Models*, The MIT Press, Cambridge, Massachusetts.

[27] Harvey, A. C. (2005). A unified approach to testing for stationarity and unit roots. *Identification and Inference for Econometric Models*, edited by D. W. K. Andrews and J. H. Stock, Cambridge University Press, New York

[28] Hills, R. G., and Trucano, T. G., (1999). Statistical Validation of Engineering and Scientific Models: Background. SAND99-1256, Sandia National Laboratories, Albuquerque, New Mexico.

[29] Hills, R. G., and Trucano, T. G., (2002). Statistical Validation of Engineering and Scientific Models: A Maximum Likelihood Based Metric. SAND2001-1783, Sandia National Laboratories, Albuquerque, New Mexico.

[30] Hills, R. G., and Trucano, T. G., (2006). Model Validation: Model Parameter and Measurement Uncertainty. *Journal of Heat Transfer*, 128, 339-351.

[31] Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. ONR Research Memorandum, Carnigie Institute 52.

[32] Hurvich, C.M. and Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.

[33] Hyndman, R. J., Koehler, A. B., Snyder, R. D., and Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18, 439-454.

[34] Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2005). Prediction intervals for exponential smoothing using two new classes of state space models. *Journal of Forecasting*, 24, 17-37.

[35] Ingolfsson, A. and Sachs, E. (1993). Stability and sensitivity of an EWMA controller. *Journal of Quality Technology*, 25(4), 271-287.

[36] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82, 34-45.

[37] Kennedy, M. C. and O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximation are available. *Biometrika*, 87(1), 1-13

[38] Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society (B)*, 63(3), 425-464

[39] Lange, K. (1999). *Numerical Analysis for Statistician*. Springer.

[40] Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Springer.

[41] Makridakis, S. and Hibon, M. (1979). Accuracy of forecasting: an empirical investigation (with discussion). *Journal of the Royal Statistical Society (A)*, 142, 97-145.

[42] Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, M., Newton, J., Parzen, E., and Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting*, 1, 111-153.

[43] Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Ord, K., Mills, T., and T. F. Simmons (1993). The M2-Competition: a real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9, 5-23.

[44] Makridakis, S. and Hibon, M. (2000). The M3-Competition: results, conclusion, and implications. *International Journal of Forecasting*, 16, 451-476.

[45] Mentzer, J. T. and Kahn, K. B. (1995). Forecasting technique familiarity, satisfaction, usage, and application. *Journal of Forecasting*, 14, 465-467.

[46] Meinnhold, R. J. and Singpurwalla, N. D. (1983). Understanding the Kalman filter. *The American Statistician*, 37(2), 123-127.

[47] Montgomery, D., Johnson, L., and Gardiner, J. (1990). *Forecasting and Time Series Analysis*. 2nd Edition, Mcgraw-Hill.

[48] Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55, 299-306.

[49] Oberkampf, W. L. and Trucano, T. G. (2000). Validation Methodology in Computational Fluid Dynamics. *Fluids 2000 Conference*, AIAA 2000-2549, Denver, Colorado.

[50] Oberkampf, W. L. and Barone, M. F. (2004). Measures of Agreement Between Computation and Experiment: Validation Metrics. *34th Fluid Dynamics Conference and Exhibit*, AIAA-2004-2626, Portland, Oregon.

[51] Ord, J. K., Koehler, A. B., and Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, 92, 1621-1629.

[52] Pegels, C. C. (1969). Exponential forecasting: some new variations. *Management Science*, 15, 311-315.

[53] Prest, A. R. (1949). Some experiments in demand analysis. *Review of Economics and Statistics*, 31, 33-49.

[54] Qian, Z. and Wu, C. F. (2005). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. School of Industrial and Systems Engineering, Georgia Institute of Technology.

[55] Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Wu, C. F. (2006). Building surrogate models based on detailed and approximate simulations. *ASME Journal of Mechanical Design*, 128, 668-677

[56] Roberts, S. A. (1982). A general class of Holt-Winters type forecasting models. *Management Science*, 28, 808-820.

[57] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and Analysis of Computer Experiments. *Statistical Science*, 4(4), 409-435.

[58] Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer.

[59] Snyder, R. D. (1985). Recursive estimation of dynamic linear models. *Journal of the Royal Statistical Society (B)*, 47, 272-276.

[60] Snyder, R. D. (2004). A pedants approach to exponential smoothing. Department of Econometrics and Business Statistics, Monash Univeristy, Australia

[61] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.

[62] Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, 19, 715-725.

[63] Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992). Screening, Predicting, and Computer Experiments. *Technometrics*, 34(1), 15-25.

[64] Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6, 324-342.

[65] Young, P. (1984). *Recursive Estimation and Time Series Analysis*. Springer-Verlag

VITA

Shuchun Wang was born in Linyi, Shanxi Province, China. She received a B.S. degree in 1996 and an M.S. degree in 1999 both in Civil Engineering from Tsinghua University, Beijing, China. She obtained an M.S. degree in Civil Engineering from the Georgia Institute of Technology in 2001. From 2001 to 2006, she was a graduate research assistant in the School of Industrial and Systems Engineering at the Georgia Institute of Technology. She obtained an M.S. degree in Statistics in 2005 and will receive a Ph.D. degree in Statistics in 2006 from the Georgia Institute of Technology.